https://doi.org/10.5194/egusphere-2026-4074
Preprint. Discussion started: 20 August 2026

Bayesian evaluation of deep learning architectures and sensor modalities for remote sensing-based driftwood segmentation in the Mackenzie Delta, Arctic Canada

Carl Stadie[1,2], Ingmar Nitze[1], Begüm Demir[2], Martin Stefan Brandt[3], and Guido Grosse[1,4]

[1]Alfred Wegener Institute Helmholtz Centre for Polar and Marine Science, 14473 Potsdam, Germany
[2]Faculty of Electrical Engineering and Computer Science, Technische Universität Berlin, 10587 Berlin, Germany
[3]Department of Geosciences and Natural Resource Management, University of Copenhagen, Copenhagen, Denmark
[4]Institute of Geosciences, University of Potsdam, 14469 Potsdam, Germany

Correspondence: Carl Stadie (carl.stadie@awi.de)

Abstract. Automated mapping of driftwood deposits along Arctic coastlines is a challenging task due to spectral ambiguity, strong depositional heterogeneity, and limited training data. Systematic comparisons of sensor modality and deep learning architecture choices remain absent for this application, leaving practitioners without evidence-based guidance on which combination to deploy. Here we present a statistical evaluation of three architectures, U-Net, Swin-U-Net, and the TerraMind foundation model, across aerial (15 cm), PlanetScope (3 m), and Sentinel-2 (10 m) imagery acquired over 10 target areas in the Mackenzie Delta, Arctic Canada. Each combination was trained 10 times and evaluated within a Bayesian hierarchical framework to account for run-to-run variability inherent to stochastic training. Choice of the sensor is the dominant performance driver, having an effect approximately four times larger than the choice of architecture in Intersection over Union (IoU) and a total sensor spread of 0.40 IoU between aerial and Sentinel-2 imagery. Architecture choice is of limited practical consequence at sub-metre and intermediate resolution, but becomes a first-order concern when constrained to coarse imagery: U-Net performs poorly on Sentinel-2 with a posterior mean IoU of 0.127, while transformer-based architectures show a more gradual performance decline. Swin-U-Net paired with PlanetScope imagery is the most competitive accessible alternative to aerial acquisition, with a 94 % posterior probability of practical equivalence to the top-ranked configuration. Conventional single-run evaluation missed this combination as a practical alternative, typically placing it at ranks 4–5. Probabilistic multi-run evaluation is therefore a necessary condition for reliable model selection in spectrally ambiguous remote sensing benchmarks, and the framework presented here could be directly transferable to similar Arctic mapping targets.

1 Introduction

Driftwood constitutes a common feature along Arctic coastlines. Composed predominantly of larch and spruce logs mobilised from boreal forests and transported via river systems to the Arctic Ocean (Alix, 2005; Murphy et al., 2021; Wohl et al., 2026), deposits range from scattered individual logs to elevated berms and driftwood mats covering up to 11–15 ha (Sendrowski et al., 2023; Stadie et al., 2025). In the otherwise treeless Arctic tundra, they provide habitat and nutrients (Grilliot et al., 2019), stabilise coastlines (Davidson et al., 2020), supply local communities with a critical wood resource (Alix, 2005; Murphy et al.,

https://doi.org/10.5194/egusphere-2026-4074
Preprint. Discussion started: 20 August 2026

2020), and represent a large carbon pool (Büntgen et al., 2026; Sendrowski et al., 2023). Their location and deposition history also serve as indicators of past environmental conditions, including sea-ice extent, storm surges, and Arctic current dynamics (Dalaiden et al., 2018; MacLeod and Dallimore, 2021). Understanding and mapping their spatial distribution is therefore essential for ecological assessments and reconstructions of environmental change (Murphy et al., 2021).

Still, automated mapping of driftwood remains difficult. Deposits are spectrally similar to bare sandy surfaces abundant in river, delta and coastal settings, their boundaries are often diffuse showing near identical responses across the red, green and blue spectrum (Dietenberger et al., 2025; Sendrowski and Wohl, 2021), and the scarcity of annotated training data limit the application of modern deep learning architectures. Thus, mapping so far relied on manual digitisation of aerial imagery, which is limited in spatial coverage due to acquisition costs and labour intensity (Husson et al., 2016; Sendrowski et al., 2023). Existing automated approaches used random forests or support vector machines (Buscombe et al., 2024; Ruiz Villanueva et al., 2024; Sendrowski and Wohl, 2021), while more recent work has adopted convolutional neural networks, in particular the U-Net and its derivatives (Sendrowski et al., 2023; Stadie et al., 2025), which have become the standard baseline for semantic segmentation in Earth observation.

The focus has likewise shifted from aerial photographs with 5–20 cm spatial resolution (Buscombe et al., 2024) to very high-resolution satellite data and, most recently, 3 m PlanetScope imagery (Sendrowski et al., 2023; Stadie et al., 2025). In parallel, transformer-based architectures (Cao et al., 2021; Liu et al., 2022) and large-scale foundation models pre-trained on vast amounts of remote sensing data (Jakubik et al., 2023, 2025; Lu et al., 2025) have achieved state-of-the-art performance in broader Earth observation benchmarks. Despite this observed shift towards automated satellite-based mapping and the emergence of more powerful deep learning architectures, systematic comparisons and subsequent application of these approaches in terms of both model architecture and input imagery remain scarce. This gap is particularly critical for driftwood mapping, where spectral and textural ambiguity and limited training data amplify variability in model performance and make results highly sensitive to methodological choices (Bouthillier et al., 2021; Henderson et al., 2018).

The decision on which sensor–architecture configuration to deploy requires reliable performance estimates. However, as neural network training is an inherently stochastic process, outcomes vary across runs due to random weight initialisation, data sampling, and augmentation (Bouthillier et al., 2021) meaning that single-run performance estimates cannot reliably represent the true capability of an architecture (Henderson et al., 2018; Reimers and Gurevych, 2018). In the driftwood mapping setting, this variability is further amplified by the properties of the target. Spectral ambiguity between driftwood and background increases sensitivity to initialisation, limited training data amplify the influence of stochastic data sampling, and strong class imbalance makes augmentation choices a substantial source of run-to-run variation (Buda et al., 2018).

Accordingly, evaluating models based on single-run metrics masks uncertainty in model ranking and limits the robustness and reproducibility of reported results. It prevents practitioners from assessing how transferable findings and expected model performance are across regions, sensors, or model configurations (Henderson et al., 2018; Kattenborn et al., 2022). To obtain meaningful and comparable performance estimates and subsequent guidance on sensor and architecture selection, evaluation frameworks must explicitly account for run-to-run variability and quantify uncertainty in model comparisons (Bouthillier et al., 2021; Maier-Hein et al., 2018; Reinke et al., 2024).

https://doi.org/10.5194/egusphere-2026-4074
Preprint. Discussion started: 20 August 2026

To address both challenges, we combine a systematic comparison of sensor–architecture combinations with a statistical, uncertainty-aware evaluation framework. We construct a multi-sensor, multi-resolution dataset of aerial (10–15 cm), PlanetScope (3 m), and Sentinel-2 (10 m) imagery acquired over 10 target areas in the Mackenzie Delta, Canada, and evaluate the U-Net (Ronneberger et al., 2015), Swin-U-Net (Cao et al., 2021), and a fine-tuned TerraMind foundation model (Jakubik et al., 2025) across all nine sensor–architecture combinations. Each combination is trained 10 times with independent initialisations to capture stochastic variability. Performance is then assessed within a Bayesian hierarchical framework that yields posterior distributions of segmentation metrics, rank probabilities, and probabilities of superiority and practical equivalence (Benavoli et al., 2017; Corani et al., 2017; Kruschke, 2018).

Our results provide a systematic, statistically grounded guidance on sensor and architecture selection for Arctic driftwood mapping, showing that sensor modality is the dominant performance driver and that Swin-U-Net paired with PlanetScope imagery is the most competitive alternative accessible for the mapping of large regions compared to aerial acquisition. We further show that conventional single-run evaluation cannot reliably reach a conclusion on sensor–architecture selection for feature mapping, underscoring the need for probabilistic multi-run frameworks in spectrally ambiguous remote sensing benchmarks such as driftwood.

2 Data

2.1 Study area

The study area comprises 10 target areas distributed across the Beaufort Delta Region, including the Mackenzie Delta and its adjacent coastlines (Fig. 1a). The Mackenzie Delta is the largest delta in the North American Arctic and the second largest Arctic delta overall (Burn and Kokelj, 2009). Covering approximately 13,000 km^2 (Sendrowski et al., 2023), it forms a complex distributary network fed by the Mackenzie, Peel, and Arctic Red rivers. Together, these rivers drain an area of approximately 1.8 million km^2, constituting the third-largest Arctic drainage basin by catchment area (Woo and Thorne, 2003; Piliouras and Rowland, 2020; Sendrowski et al., 2023).

The delta receives large quantities of driftwood that are transported downstream towards the Beaufort Sea. Most of this material originates from the boreal forested tributaries of the Mackenzie River, particularly the Liard River catchment (Sendrowski et al., 2023). Although wood input is generally continuous, episodic wood-flood events periodically mobilise substantial quantities of deadwood upstream of the delta. These events, which occur with an estimated return interval of approximately five years, play a major role in redistributing wood throughout the deltaic system (Kramer et al., 2017; Sendrowski et al., 2023).

Because only a fraction of the transported wood ultimately reaches the coast, extensive driftwood accumulations occur throughout both the delta and its adjacent shorelines (Sendrowski et al., 2023). Within the delta, deposits typically consist of scattered individual logs and driftwood berms aligned along river channels (Sendrowski et al., 2023; Sendrowski and Wohl, 2021). Along the outer coastline, however, driftwood frequently accumulates in large mats that may extend over several hectares within sheltered embayments (Stadie et al., 2025).

https://doi.org/10.5194/egusphere-2026-4074
Preprint. Discussion started: 20 August 2026

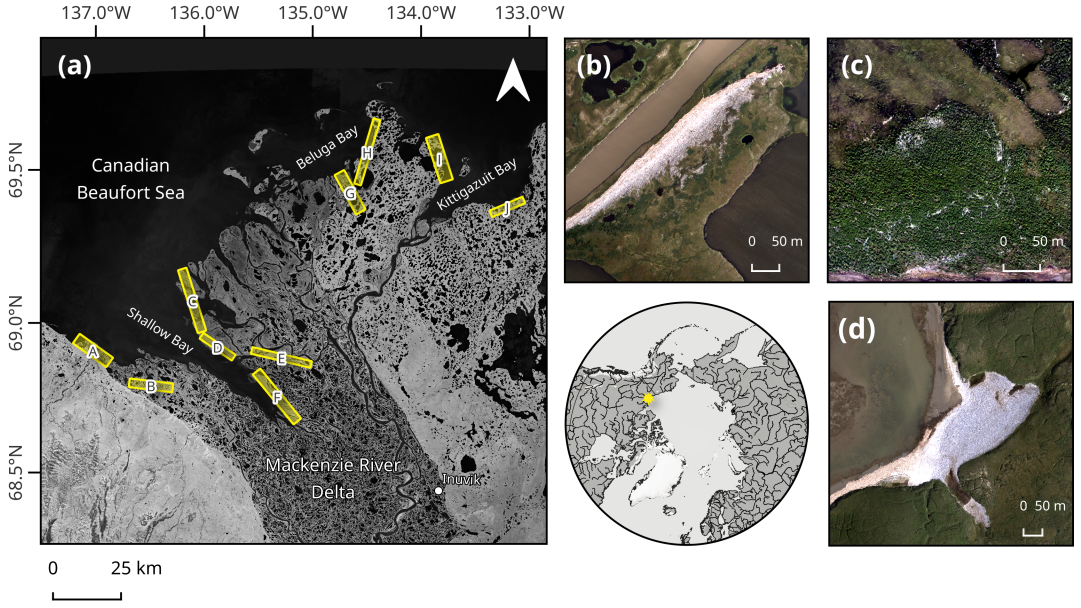

Figure 1. Study area and representative driftwood deposits. (a) Location of the 10 target areas (A–J) within the Mackenzie Delta and adjacent Beaufort Sea coastline, Canada (see Table S1 for details). (b) Driftwood mat in target area A surrounded by smaller berms and scattered logs. (c) Vegetation-overgrown driftwood deposits in target area D, consisting of small mats and individual logs, representative of typical deltaic deposits. (d) Driftwood mat within a coastal embayment in target area H. Wood accumulates against elevated terrain, creating a distinct landward boundary, whereas the seaward margin is less clearly defined and interspersed with sediment.

The 10 target areas were selected from locations previously identified as exhibiting particularly high driftwood occurrence (Sendrowski et al., 2023; Stadie et al., 2025). While site selection focused on maximising wood presence, the resulting dataset captures substantial variability in deposit density, morphology, and vegetation cover representative of driftwood deposits across the Mackenzie Delta.

Areas A and B, located along the western shoreline of the Shallow Bay channel, are characterised by extensive driftwood mats interspersed with berms and scattered logs (Fig. 1b). On the eastern side of Shallow Bay, areas C–F predominantly contain typical deltaic deposits consisting of driftwood berms and scattered logs, many of which are partially or fully overgrown by vegetation (Fig. 1c). In the eastern part of the delta, areas G–J encompass a diverse range of coastal deposits, from isolated logs on beaches, floodplains, and sandbanks to large, spatially distinct driftwood mats occupying protected coastal embayments (Fig. 1d).

2.2 MACS aerial imagery

For each of the 10 target areas, aerial images were acquired during the Perma-X 2025 flight campaign on board the Alfred Wegener Institute's (AWI) Polar-5 aircraft (Wesche et al., 2016) equipped with the Modular Aerial Camera System (MACS)

https://doi.org/10.5194/egusphere-2026-4074
Preprint. Discussion started: 20 August 2026

developed by the German Aerospace Center (DLR) in its MACS POLAR-24 configuration (Grosse et al., 2025; Lehmann et al., 2011). The MACS is highly suitable for use in polar regions, including challenging illumination conditions and strong contrasts between dark surfaces (e.g. exposed soils or water) and bright surfaces (e.g. snowfields, bare sand, or patches of bleached driftwood) in the same area by using dedicated acquisition/processing strategies tailored to these environments (Rettelbach et al., 2024).

All flights were conducted between 21 July and 12 August 2025 at an altitude of approximately 850 m above ground level, resulting in a ground sampling distance (GSD) of 15 cm (Grosse et al., 2025; Rettelbach et al., 2024). This resolution is comparable to, or higher than ground resolutions commonly used in driftwood mapping studies (Sendrowski et al., 2023; Sendrowski and Wohl, 2021; Stadie et al., 2025). Because the aerial images provided the highest spatial detail, they served as the reference dataset for our analysis. A detailed overview of the image acquisitions is provided in Supplementary Table S1.

The MACS imagery consists of four spectral bands (blue, green, red, near-infrared). For our 10 study areas, a total of 56,920 images were processed into orthomosaics using a structure-from-motion workflow (Rettelbach et al., 2024). In total, the mosaics cover an area of 615 km^2.

2.3 PlanetScope imagery

In addition to the aerial imagery, we used high-resolution images from the PlanetScope nanosatellite constellation operated by Planet Labs PBC, which provides near-daily coverage at approximately 3 m spatial resolution with four multispectral bands: blue, green, red, and near-infrared (Planet Labs PBC, 2023). PlanetScope has been used previously for large-scale driftwood mapping because it offers a practical compromise between spatial detail and regional coverage. Although individual logs are often smaller than a 3 m pixel, larger accumulations can still be captured with useful fidelity, while total driftwood area may be underestimated relative to sub-metre aerial reference data (Stadie et al., 2025).

Data were provided under a research licence via the Planet API using the PSScene analytic_sr bundle, which provides atmospherically corrected surface reflectance. To minimise temporal mismatch and differences in ground conditions, we selected cloud-free images acquired closest to the aerial survey dates. The mean absolute acquisition-time difference was 1.1 days. Further details on acquisition time are provided in Supplementary Fig. S2.

2.4 Sentinel-2 imagery

Sentinel-2 imagery was included to characterise model behaviour using the highest spatial resolution available from a freely accessible satellite mission, and to provide a direct evaluation of TerraMind in its native training domain (Jakubik et al., 2025). Sentinel-2 offers global coverage at 10-20 m spatial resolution, depending on wavelength (Drusch et al., 2012). Although its moderate resolution limits the detection of individual logs and small deposits, it provides a cost-free and easily accessible alternative to PlanetScope or aerial data. It is also widely used as the basis for foundation models, which can be advantageous in tasks such as driftwood mapping where training data are limited (Cong et al., 2022; Jakubik et al., 2025; Mañas et al., 2021).

We used 10 bands spanning the visible, near-infrared, and shortwave-infrared regions; 20 m bands were resampled to 10 m to align with the overall image stack. As with the PlanetScope imagery, scenes were selected based on temporal proximity to

https://doi.org/10.5194/egusphere-2026-4074
Preprint. Discussion started: 20 August 2026

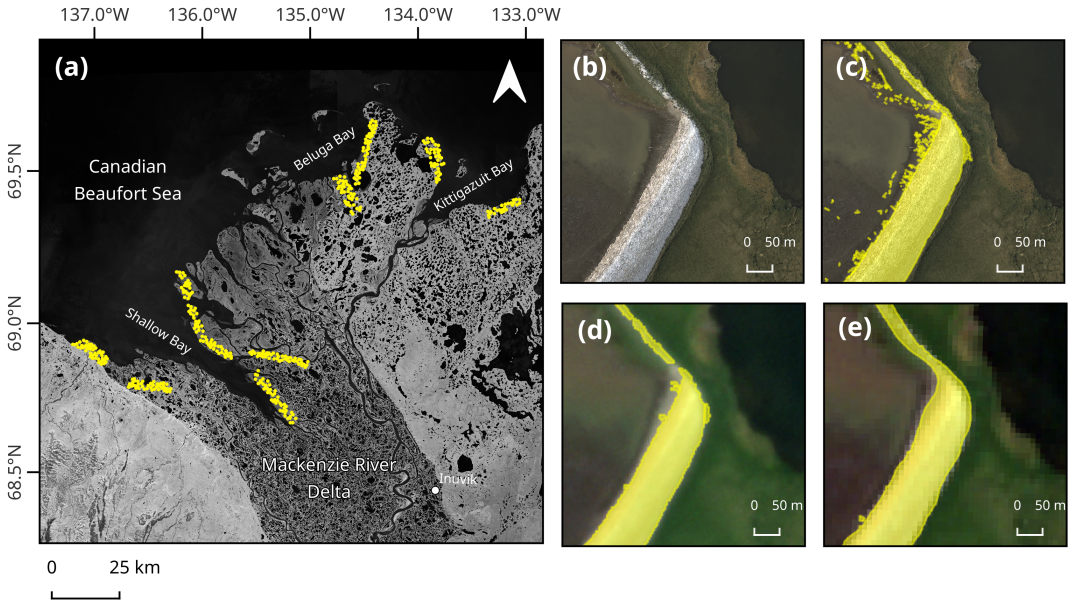

Figure 2. Training tile distribution and sensor-specific annotations. (a) Spatial distribution of the 400 annotated training tiles (yellow dots) across the 10 target areas in the Mackenzie Delta, Canada, displayed on a Sentinel-1 synthetic aperture radar (SAR) backscatter mosaic. (b–e) Example driftwood deposit represented at different sensor resolutions: (b) aerial imagery (15 cm resolution) without annotations; (c) aerial imagery with driftwood labels (yellow); (d) PlanetScope imagery (3 m resolution) with corresponding labels; and (e) Sentinel-2 imagery (10 m resolution) with corresponding labels. Labels were generated independently for each sensor at its native spatial resolution rather than derived from a common reference dataset. Consequently, the number, size, and boundary delineation of mapped driftwood deposits vary among sensor modalities, reflecting differences in spatial resolution and observable detail.

the aerial acquisition dates and low cloud cover. The mean absolute temporal offset was 2.4 days. Additional information on scene acquisition time is given in Supplementary Fig. S2.

3 Sensor-specific annotation and dataset construction

We constructed a multi-sensor dataset by annotating 400 tiles of 500×500 m each, covering a combined area of 177 km^2. Tiles were distributed approximately evenly across the 10 target areas, ensuring spatial coverage of each depositional setting. This area-stratified approach was chosen over random sampling to maintain a reasonable foreground-background balance across the dataset as random sampling across the full study area would have yielded a training set dominated by background pixels, particularly at coarser resolutions where wood occupies a very small fraction of any given tile. In total, 28.8 % of targeted areas are incorporated into annotated tiles. The extent of the training tiles is shown in Fig. 2.

The tiles span the full range of depositional settings encountered in the study area, including extensive driftwood mats, berms, and scattered logs, as well as spectrally or texturally similar non-wood surfaces such as beaches, sandbars, floodplains,

https://doi.org/10.5194/egusphere-2026-4074
Preprint. Discussion started: 20 August 2026

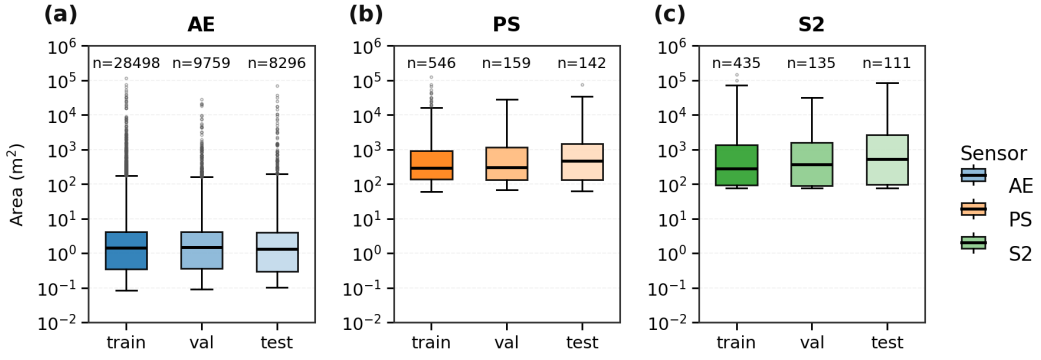

Figure 3. Distribution of driftwood object areas across the training, validation, and test datasets for each sensor. Boxes indicate the interquartile range (IQR), with the horizontal line marking the median. Whiskers extend to 1.5 × IQR, and grey circles denote outliers. The y-axis is displayed on a logarithmic scale to accommodate the wide range of object sizes. Sample sizes (*n*) indicate the number of individual driftwood objects in each dataset split. The substantially larger number of objects in the aerial imagery dataset (a) reflects its finer spatial resolution, which resolves individual driftwood pieces that are merged into larger polygons or become undetectable at the coarser PlanetScope (b) and Sentinel-2 (c) resolutions. Within each panel, dark, medium, and light shading correspond to the training, validation, and test datasets, respectively. Details of the data partitioning procedure are provided in Sect. 4.2.

vegetated areas, river channels, and open water. For each image modality, annotations were created independently, rather than by resampling a single label set. To reduce subjectivity and improve consistency during labelling, we used the GeoSAM plug-in for QGIS (Zhao et al., 2023) to generate initial polygon proposals, which were subsequently refined or manually digitised where automated segmentation failed to delineate driftwood accurately. Individual logs were labelled as separate entities, whereas densely packed accumulations like berms or mats were labelled as single deposits.

In the aerial imagery, a total of 46,553 deposits were delineated, covering 1.9 km^2. In the PlanetScope data, 847 deposits amounting to 1.86 km^2 were mapped. For Sentinel-2, the coarser spatial resolution resulted in 681 labelled deposits with a combined area of 1.91 km^2. As expected, differences in spatial resolution influenced both the number and size of labelled entities (Fig. 3). Deposits labelled in the aerial imagery exhibit an average area of 41.01 m^2, compared with 2196.34 m^2 for PlanetScope and 3221.54 m^2 for Sentinel-2. Maximum deposit areas are broadly comparable across sensors (aerial: 124,907 m^2, PlanetScope: 125,257 m^2, Sentinel-2: 149,623 m^2), indicating that the large differences in the average area are driven by the lower end of the size distribution, where small deposits and individual logs tend to fall below the detection threshold at coarser resolutions.

Resolution-driven detectability also resulted in differing proportions of annotated tiles across modalities: For the aerial images, 72.25 % of annotation areas contain driftwood labels, compared with 44.5 % for PlanetScope and only 42 % for Sentinel-2. The labelled annotation areas were rasterized for further processing, matching the spatial resolution of the respective input data.

https://doi.org/10.5194/egusphere-2026-4074
Preprint. Discussion started: 20 August 2026

4 Methods

4.1 Architectures

We chose three deep learning architectures to be trained on each image modality: U-Net, Swin-U-Net, and TerraMind. While these form just a small subset of the extensive and growing pool of deep learning architectures, they represent three evolutionary stages: convolutional baselines, transformer-based designs, and large pretrained foundation models.

4.1.1 U-Net

Since its introduction in 2015, U-Net has become a widely used architecture for segmentation tasks in Earth Observation (Zhu et al., 2017), including driftwood mapping. Despite the development of derivatives, the original U-Net remains a common baseline against which new architectures are evaluated (Zhu et al., 2017).

Our implementation follows the canonical encoder-decoder design with four downsampling and four upsampling stages (Ronneberger et al., 2015). Each stage comprises two 3×3 convolutions followed by batch normalisation and ReLU activations Agarap (2018) while skip connections concatenate encoder and decoder feature maps at matching spatial resolutions (Ronneberger et al., 2015). A final 1×1 convolution with sigmoid activation produces per-pixel driftwood probabilities.

Relative to the original specification, we introduce several adaptations to address the class imbalance and scale characteristics of driftwood deposits. First, we use Tversky loss (Salehi et al., 2017), which is well suited to highly unbalanced foreground-background segmentation problems. Second, we use dilated convolutions (Chen et al., 2018) to increase the effective receptive field without increasing the number of parameters. We further enabled dropout to capture aleatoric and epistemic uncertainty.

4.1.2 Swin-U-Net

Transformer-based architectures have emerged as strong competitors to purely convolutional networks across a range of computer vision benchmarks. Vision Transformers (ViTs) operate on images decomposed into fixed-size patches and apply self-attention to model long-range dependencies, thereby capturing global relationships across the image (Dosovitskiy et al., 2021). Building on these ideas, Swin-U-Net adapts the U-shaped encoder-decoder design by replacing convolutional blocks with hierarchical Swin-Transformer blocks, which reduce the computational cost of self-attention by operating within local shifted windows rather than globally (Cao et al., 2021; Liu et al., 2022).

Inside Swin-U-Net, the encoder partitions the input into non-overlapping windows and applies windowed multi-head self-attention (Cao et al., 2021; Liu et al., 2022). Successive stages merge patches into coarser token representations, forming multi-scale features at resolutions from 1/4 to 1/32 of the input. The decoder mirrors this hierarchy using patch-expansion layers and skip connections to progressively recover spatial resolution (Cao et al., 2021).

As transformer models typically benefit from large-scale pretraining, we initialise Swin-U-Net using Swin-T weights pretrained on ImageNet-1k (Deng et al., 2009; Liu et al., 2022). Because these weights are defined for three-channel RGB inputs, for subsequent training, we initialise weights for additional spectral bands by filling them with the mean of the learned RGB

https://doi.org/10.5194/egusphere-2026-4074
Preprint. Discussion started: 20 August 2026

filters. Similar to the U-Net we used Tversky loss (Salehi et al., 2017) to represent the class imbalance between driftwood and background in the loss function. Similar to the U-Net, dropout was enabled.

4.1.3 TerraMind

Recently, foundation models have gained increasing popularity, not only in general computer vision but also in Earth observation. Foundation models in Earth observation are large networks pre-trained on extensive image collections to learn generic spectral-textural representations that can be adapted to different downstream tasks with relatively limited labelled data (Bommasani et al., 2022). TerraMind is a state-of-the-art transformer-based encoder pre-trained on global Sentinel-2/Sentinel-1 optical and synthetic aperture radar satellite imagery and derived data products like land cover classifications, and elevation models (Jakubik et al., 2025). Such pretraining is expected to improve data efficiency and provide a degree of robustness to variations in illumination, background texture, and geomorphic setting (Jakubik et al., 2025; Bommasani et al., 2022).

To use TerraMind for driftwood mapping, we build on the TerraTorch implementation and instantiate the TerraMind v1 base backbone coupled with a UPerNet decoder for semantic segmentation (Gomes et al., 2025; Xiao et al., 2018). The backbone operates on Sentinel-2 L2A embeddings; for 4-band PlanetScope and aerial imagery we map the input channels (blue, green, red, NIR) to the corresponding Sentinel-2 bands (BLUE, GREEN, RED, NIR_NARROW), reusing the same pre-trained weights across sensors. As with the aforementioned architectures, the loss function used was Tversky loss.

For uncertainty quantification, dropout is applied only within the UPerNet decoder head during sampling (Gal and Ghahramani, 2016). The TerraMind backbone weights are frozen throughout fine tuning and only unfrozen every n epochs (as determined during hyperparameter tuning). Consequently, epistemic uncertainty estimates for TerraMind configurations mostly reflect parameter uncertainty in the decoder only. This is not a measurement artefact but a structural property of foundation model fine-tuning. The frozen encoder produces consistent representations across runs by design, thus predictions are highly reproducible regardless of initialisation.

4.2 Experimental design

To assess the performance of each model-architecture combination, the annotation areas for each image modality were split into training, validation, and test sets using a stratified partitioning strategy to preserve class prevalence across subsets (Kohavi, 1995; Pawłuszek-Filipiak and Borkowski, 2020). In total, 60 % of the annotated tiles were used for training and 20 % for validation during training, while the remaining 20 % were withheld as an independent test set for subsequent model comparison. For each image modality, the assignment of tiles to the three subsets was kept identical across all architectures. Data-splits were performed at the tile level, with annotation tiles assigned to subsets stratified by driftwood area to ensure comparable class prevalence across training, validation, and test sets. The same tile assignments were used across all three sensors and all architectures to ensure that performance differences reflect model behaviour rather than differences in data exposure.

Given the strong influence of hyperparameter choice on model performance, we carried out a dedicated hyperparameter search for every architecture-image combination. For all combinations, the search space included the choice of optimiser, learning rate and weight decay. Furthermore, the use of a learning rate scheduler and batch-specific parameters such as the

https://doi.org/10.5194/egusphere-2026-4074
Preprint. Discussion started: 20 August 2026

proportion of positive frames within a batch and the minimum fraction of positive pixels required for a frame to be included in the training loop were optimised. In addition, we explored values for the parametrisation of the loss function.

We further optimised model-specific hyperparameters. For the U-Net, these consisted of the dilation rate, number of feature channels, L2 regularisation weight, and dropout probability. For the Swin-U-Net, we tuned the base channel dimension, patch and window size, and stochastic depth. For the TerraMind-based configurations, we optimised, among others, the decoder type, decoder head size, and backbone learning rate. A detailed overview of all hyperparameters and their respective search ranges is provided in Supplementary Table S3.

Hyperparameters were optimised in two stages. In the first stage, we used Hyperband optimisation (Li et al., 2018) to conduct a broad search of the joint hyperparameter space and to identify promising configurations. In a second stage, we applied Bayesian optimization (Bergstra et al., 2011; Snoek et al., 2012) to further refine continuous hyperparameters while keeping the discrete choices obtained from the Hyperband stage fixed.

After hyperparameter optimization, each model-image combination was trained on the corresponding training dataset using its best-performing configuration. During training, patches were randomly sampled from the training frames using the sampling scheme determined in the hyperparameter search. This ensured that all models were exposed to a comparable mixture of foreground and background examples despite the strong class imbalance (Buda et al., 2018; Pala et al., 2023). A detailed overview of the resulting model parameterisations is provided in Supplementary Table S4.

For each combination, we performed 10 independent training runs differing only in random weight initialisation and stochastic data sampling (Bouthillier et al., 2021; Reimers and Gurevych, 2018). Models were trained for 100 epochs with 500 optimisation steps per epoch using Tversky loss (Salehi et al., 2017) and the optimiser-scheduler pair selected during hyperparameter optimisation. The checkpoint with the lowest validation loss was retained for evaluation.

During evaluation, each trained model was applied to the withheld test set and produced per-pixel probability maps, which were binarized using a fixed threshold of 0.5. We report Intersection over Union (IoU) as our primary overlap metric, defined in Eq. (1), as it provides a strict, widely adopted measure of spatial agreement that directly penalizes false positives and false negatives (Jaccard, 1912):

$$\mathrm{IoU}(A,B) = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

where A denotes the set of foreground pixels in the binarized model prediction and B denotes the corresponding set of foreground pixels in the ground truth mask. We additionally report the Dice Similarity Coefficient (DSC) (Dice, 1945), defined in Eq. (2), as it is the most commonly used overlap score in segmentation studies in addition to IoU and provides an easily interpretable, comparable summary of overall segmentation quality:

$$\mathrm{DSC}(A,B) = \frac{2|A \cap B|}{|A| + |B|} = \frac{2 \cdot \mathrm{PPV} \cdot \mathrm{Sensitivity}}{\mathrm{PPV} + \mathrm{Sensitivity}} \tag{2}$$

where PPV (positive predictive value) and Sensitivity (true positive rate) denote precision and recall, respectively, computed pixel-wise from A and B as defined above. To specifically assess boundary quality while accounting for annotation imprecision

https://doi.org/10.5194/egusphere-2026-4074
Preprint. Discussion started: 20 August 2026

(Maier-Hein et al., 2024), we report the Normalised Surface Distance (NSD) (Nikolov et al., 2021), defined in Eq. (3), where S_A and S_B denote the boundaries of A and B, respectively, and $\mathcal{B}_A^{(\tau)}$ and $\mathcal{B}_B^{(\tau)}$ denote the boundary regions within tolerance τ around the boundaries of A and B, respectively:

$$\mathrm{NSD}(A,B)^{(\tau)} = \frac{\left|S_A \cap \mathcal{B}_B^{(\tau)}\right| + \left|S_B \cap \mathcal{B}_A^{(\tau)}\right|}{|S_A| + |S_B|} \tag{3}$$

For NSD calculation, τ was set to one pixel, making the tolerance sensor-relative rather than absolute. One pixel corresponds to approximately 10–15 cm in the aerial imagery, 3 m in PlanetScope, and 10 m in Sentinel-2. This choice reflects a fundamental constraint of cross-resolution boundary evaluation for which no universally fair solution exists. Setting τ to a finite absolute value matching the aerial pixel size would render NSD unmeasurable at Sentinel-2's 10 m resolution, while setting it to match Sentinel-2's resolution would grant aerial models a tolerance of approximately 66 pixels, neutralising the boundary precision advantage of high-resolution imagery. The sensor-relative one-pixel tolerance therefore represents the least-biased available choice, ensuring that NSD captures how well a model delineates deposits given the information available in that modality, and prevents penalising coarser-resolution predictions for failing to capture sub-pixel boundaries that the imagery cannot represent (Nikolov et al., 2021; Ostmeier et al., 2023).

In addition to overlap and boundary metrics, predictive uncertainty was quantified via Monte Carlo stochastic inference with stochastic dropout active at inference time (Gal and Ghahramani, 2016). For a pixel with stochastic probability samples $\{p_t\}_{t=1}^{T}$, we estimated aleatoric uncertainty as the expected Bernoulli variance $\mathbb{E}[p_t(1-p_t)]$, reflecting irreducible data ambiguity, and epistemic uncertainty as the variance of the mean prediction $\mathrm{Var}(p_t)$, reflecting model parameter uncertainty (Kendall and Gal, 2017).

4.3 Bayesian model comparison

Instead of comparing models using single summary statistics, we explicitly model the variability across repeated training runs. This allows us to estimate not only which model performs best on average, but also how certain we are about these differences. For this, we used a Bayesian hierarchical model to assess model performance across metrics while accounting for run-to-run variability caused by the inherent stochasticity of model training (Bouthillier et al., 2021). An advantage of Bayesian hierarchical models is that they provide a statistical framework that allows information to be shared across groups, producing more stable estimates than analysing each group independently.

For each performance metric, a hierarchical group model was fitted in which group-level means are drawn from a shared population distribution, providing modest regularisation that stabilises group-specific estimates (Benavoli et al., 2017; Corani et al., 2017; Gelman, 2005, 2006). Here, group-level means represent the estimated true average performance per configuration, as opposed to the observed score from any single training run. Through partial pooling, estimates for each configuration are regularised towards a shared average across all configurations, while variability within each configuration is also estimated hierarchically through a shared half-normal hyperprior, which allows the model to borrow variance information across sensor-architecture combinations rather than estimating each group's dispersion independently. A hyperprior is a prior placed on a

https://doi.org/10.5194/egusphere-2026-4074
Preprint. Discussion started: 20 August 2026

parameter that itself governs other parameters; for instance, a hyperprior on the spread of group means controls how far each configuration's estimated average may deviate from the overall average. With this approach, estimates for each configuration are gently regularised toward a shared average, reducing the risk of overinterpreting extreme results from small sample sizes.

Observed metric values were modelled on a transformed scale appropriate to their distributional support. Overlap metrics were modelled on the logit scale, to ensure valid Gaussian modelling, as is standard for proportion-like data (**?**). Uncertainty metrics, which are strictly positive and unbounded above, were modelled on the log scale to stabilise skewed positive values. The intercept hyperprior μ_0 was centred on the median of the transformed observations to provide a data-informed but weakly regularising prior. Group means were assigned partially pooled priors ($\mu \sim \mathrm{Normal}(\mu_0, \tau)$), where τ was given a half-normal hyperprior following Gelman (2006). Observations were modelled using Gaussian likelihoods with group-specific standard deviations. A graphical overview of the model can be found in Supplementary Fig. S5.

All models were sampled using the No-U-Turn Sampler (NUTS) with 2000 draws across 4 chains following 2000 tuning iterations, targeting an acceptance rate of 0.9 to ensure adequate exploration of the posterior (Hoffman and Gelman, 2014). Model adequacy was assessed through prior and posterior predictive checks (Gabry et al., 2019), see Supplementary Figs. S6 and S7. Prior predictive checks confirmed that the chosen priors placed substantial probability mass across the plausible range of metric values without concentrating near specific outcomes, with observed group means falling comfortably within the prior 90 % highest density intervals for all configurations. Posterior predictive checks confirmed adequate model fit for overlap metrics across all sensor-architecture combinations, with the posterior predictive distribution closely recovering the multimodal structure of the observed IoU and Dice distributions and individual run values falling within the 90 % posterior predictive intervals. For uncertainty metrics, the log-scale Gaussian likelihood slightly underestimates the right tail for configurations exhibiting near-zero epistemic uncertainty with occasional high-variance runs, consistent with the structural differences in uncertainty estimation across architectures described in Sect. 4.1.

4.3.1 Factorial analysis

To differentiate the independent and interactive contributions of sensor and architecture choice (factors) to model performance, we implemented a 3×3 factorial Bayesian model with effects coding, analogous to a two-way ANOVA (Gelman, 2005; Schad et al., 2020). The model can be formulated as shown in Eq. (4):

$$\mu_{ij} = \beta_0 + \sum_k \beta_k^d D_{ki} + \sum_l \beta_l^a A_{lj} + \sum_k \sum_l \beta_{kl}^{\mathrm{int}} (D_{ki} \times A_{lj}) \tag{4}$$

where β_0 represents the grand mean, β^d and β^a are vectors of dataset and architecture main effects respectively, β^{int} captures pairwise interactions with D and A representing effects-coded design matrices (Schad et al., 2020). Main effects represent the average influence of one factor on performance, independent of the other, whereas interaction effects represent the degree to which the effect of one factor depends on the level of the other. In effects coding, each level of a factor is contrasted against a reference level using $+1$, 0, or -1 entries. We assigned weakly informative priors matching those used in the hierarchical group models to all effect parameters (Gelman, 2006). Posterior effects of architecture and image choice were computed relative to

https://doi.org/10.5194/egusphere-2026-4074
Preprint. Discussion started: 20 August 2026

the grand mean and additionally in contrast to the conventional baseline of U-Net and aerial imagery or the combination of both.

Practical significance was assessed using a region of practical equivalence (ROPE) (Kruschke, 2018). The ROPE defines a predefined threshold below which a difference between configurations is considered negligible. We set the ROPE to ± 0.05 on the logit scale for overlap metrics, corresponding to an approximately 1 % absolute difference in IoU, and to $\pm \log(1.05)$ on the log scale for uncertainty metrics, corresponding to an approximately 5 % multiplicative change.

4.3.2 Probabilistic ranking

For single-metric rankings, we used posterior draws of group-level means to simulate the full ranking distribution across all sensor-architecture combinations. For each of 8000 posterior samples, configurations were ranked from best to worst according to the directionality of the metric. From these simulated rankings, we computed the probability of each configuration achieving rank k ($\Pr(\text{rank} = k)$), its expected rank $E[\text{rank}] = \sum_k k \cdot \Pr(\text{rank} = k)$, and the probability of being the best-performing configuration (Benavoli et al., 2017; Corani et al., 2017).

For multi-metric aggregation, posterior draws of group means were first aligned to a common direction of improvement, then standardised within each posterior draw using standard z-scores. The standardised scores were combined using pre-specified weights reflecting the relative operational importance of each metric (IoU: 0.40, NSD: 0.35, Dice: 0.15, epistemic uncertainty: 0.05, aleatoric uncertainty: 0.05), yielding a composite utility score for each configuration and posterior draw (Saisana et al., 2005). Metric weights were chosen to balance overlap and boundary metrics, while incorporating aleatoric and epistemic uncertainty. Ranking probabilities and expected ranks were then derived from this composite score analogously to the single-metric case (Benavoli et al., 2017; Corani et al., 2017).

To quantify the separation between the top-ranked configuration and the runner-up, we computed two complementary summaries: the probability that the composite score difference exceeded a practical equivalence threshold on the composite scale (ROPE $= 0.15$), and the probability that the top configuration outscored the runner-up. We emphasise that multi-metric rankings depend on the chosen weighting scheme and aggregation method, and are therefore intended as illustrative decision-support summaries rather than definitive model selection criteria (Saisana et al., 2005).

4.3.3 Comparison to single run rankings

To assess whether single-run performance estimates produce reliable rankings, we compared rank orderings derived from individual training runs against the posterior expected ranks from the hierarchical model. For each of the 10 independent training runs per configuration, all nine sensor-architecture combinations were ranked by their test-set metric value, yielding 10 single-run rank vectors per metric. Exact rank agreement was computed as the proportion of configurations assigned the same rank as in the posterior expected ordering, averaged across runs. Ordinal consistency was quantified using Kendall's τ between each single-run rank vector and the posterior expected rank vector, again averaged across runs (Maier-Hein et al., 2018). This analysis was conducted for each individual metric and for the composite score, to assess whether multi-metric aggregation stabilises rank estimates relative to single-metric evaluation.

https://doi.org/10.5194/egusphere-2026-4074
Preprint. Discussion started: 20 August 2026

Table 1. Posterior mean and 95 % highest density interval (HDI) for segmentation performance (IoU, Dice, NSD) and predictive uncertainty (epistemic, aleatoric) across model architecture and sensor modality (AE: aerial imagery, PS: PlanetScope, S2: Sentinel-2).

Configuration	IoU ↑		Dice ↑		NSD ↑		$\text{Uncertainty}_{\text{epist.}}$ ↓		$\text{Uncertainty}_{\text{alea.}}$ ↓	
	mean	HDI 95 %[a]	mean	HDI 95 %	mean	HDI 95 %	mean	HDI 95 %	mean	HDI 95 %
U-Net | AE	0.798	[0.787, 0.811]	0.887	[0.879, 0.895]	0.391	[0.380, 0.399]	0.000	[0.000, 0.000]	0.000	[0.000, 0.000]
Swin-U-Net | AE	0.764	[0.737, 0.788]	0.866	[0.849, 0.883]	0.370	[0.363, 0.399]	0.003	[0.002, 0.003]	0.000	[0.000, 0.000]
TerraMind | AE	0.753	[0.747, 0.759]	0.859	[0.855, 0.863]	0.230	[0.228, 0.233]	0.000	[0.000, 0.000]	0.001	[0.001, 0.001]
U-Net | PS	0.641	[0.617, 0.666]	0.781	[0.763, 0.798]	0.482	[0.445, 0.519]	0.001	[0.001, 0.001]	0.000	[0.000, 0.000]
Swin-U-Net | PS	0.600	[0.585, 0.616]	0.750	[0.739, 0.762]	0.546	[0.536, 0.590]	0.001	[0.001, 0.001]	0.000	[0.000, 0.000]
TerraMind | PS	0.568	[0.564, 0.573]	0.725	[0.721, 0.728]	0.382	[0.366, 0.399]	0.000	[0.000, 0.000]	0.001	[0.000, 0.001]
U-Net | S2	0.127	[0.114, 0.140]	0.225	[0.204, 0.245]	0.604	[0.590, 0.620]	0.001	[0.001, 0.001]	0.000	[0.000, 0.000]
Swin-U-Net | S2	0.548	[0.533, 0.562]	0.708	[0.696, 0.719]	0.546	[0.516, 0.577]	0.001	[0.001, 0.001]	0.000	[0.000, 0.000]
TerraMind | S2	0.415	[0.404, 0.426]	0.587	[0.575, 0.597]	0.497	[0.472, 0.523]	0.000	[0.000, 0.000]	0.002	[0.001, 0.002]

[a]HDI: the range containing the 95 % most probable values of the posterior distribution, where every point inside the interval has higher probability than any point outside it.

5 Results

5.1 Overview of segmentation performance

We evaluated nine sensor-architecture combinations, each trained 10 times with identical train/validation/test splits to account for run-to-run variability arising from stochastic weight initialisation and data sampling. Performance was assessed on the withheld test set using five metrics: Intersection over Union (IoU), Dice Similarity Coefficient (Dice), Normalised Surface Distance (NSD), and mean epistemic and aleatoric uncertainty. The raw distributions across 10 independent runs for all nine combinations are shown in Fig. 4, and posterior means and 95 % HDIs derived from the hierarchical model are provided in Table 1 as a reference for all subsequent comparisons.

For the overlap metrics IoU and Dice, performance degrades monotonically with decreasing spatial resolution within every architecture, with the exception of U-Net on Sentinel-2, which exhibits a median IoU of approximately 0.13, substantially below all other configurations. NSD shows an inverted pattern, with values increasing at coarser spatial resolution. Run-to-run variability within any single combination is small relative to the inter-sensor spread, with interquartile ranges of 0.02–0.05 IoU across 10 runs. TerraMind configurations show near-zero epistemic uncertainty across all sensors, several orders of magnitude below U-Net and Swin-U-Net, accompanied by elevated aleatoric uncertainty most prominently on Sentinel-2.

5.2 Factorial decomposition of sensor and architecture effects

To decompose observed performance into independent contributions of sensor modality and architecture choice, we fitted a 3×3 factorial Bayesian model with effects coding. All effects are expressed as absolute unit changes (Δ) from the grand mean and classified relative to a region of practical equivalence (ROPE) of 1 percentage point for overlap metrics and a 5 %

https://doi.org/10.5194/egusphere-2026-4074
Preprint. Discussion started: 20 August 2026

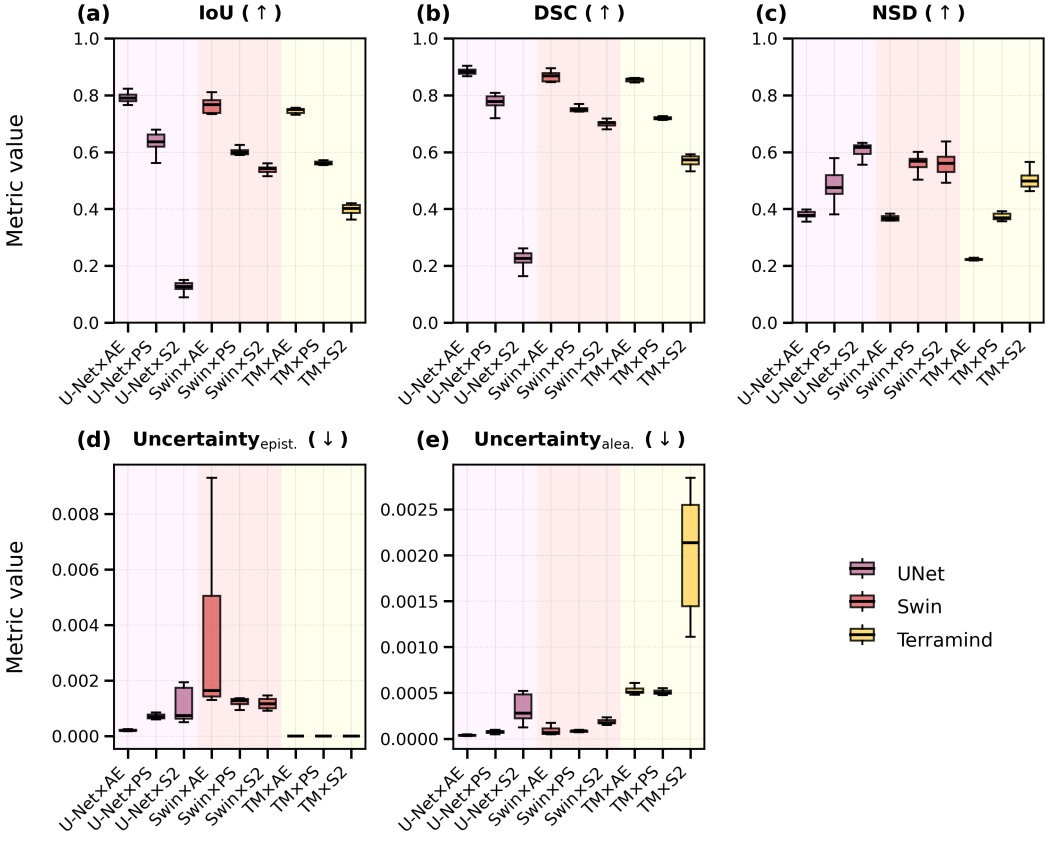

Figure 4. Boxplots of raw test-set performance across 10 independent runs for all nine sensor–architecture combinations, grouped by architecture. Panels show (a) IoU, (b) Dice similarity coefficient, (c) normalised surface distance, (d) mean epistemic uncertainty, (e) mean aleatoric uncertainty. Boxes show the interquartile range; whiskers extend to $1.5 \times$ IQR. Overlap metrics degrade monotonically with decreasing spatial resolution within every architecture, with the exception of U-Net on Sentinel-2, which exhibits a median IoU substantially below all other configurations. Normalised surface distance shows an inverted pattern, with values increasing at coarser resolution, consistent with the sensor-relative tolerance described in Sect. 4.2. TerraMind configurations show near-zero epistemic uncertainty across all sensors, accompanied by elevated aleatoric uncertainty most prominently on Sentinel-2. Note that uncertainty panels use different y-axis scales from the overlap metrics.

multiplicative change for measures of uncertainty. Posterior distributions of all effects across the five evaluation metrics are shown in Fig. 5.

For IoU and Dice, dataset main effects are large and decisively separated from zero and from the ROPE. Aerial imagery produces a strong positive main effect (+0.19 IoU; $\mathrm{HDI}_{95\,\%}$ [+0.19, +0.20]) and Sentinel-2 a comparably large negative effect (-0.22 IoU; $\mathrm{HDI}_{95\,\%}$ [-0.22, -0.21]), with PlanetScope falling near the grand mean (+0.02 IoU). The total sensor spread of approximately 0.40 IoU is roughly four times larger than the largest architecture main effect observed for these metrics. For

https://doi.org/10.5194/egusphere-2026-4074
Preprint. Discussion started: 20 August 2026

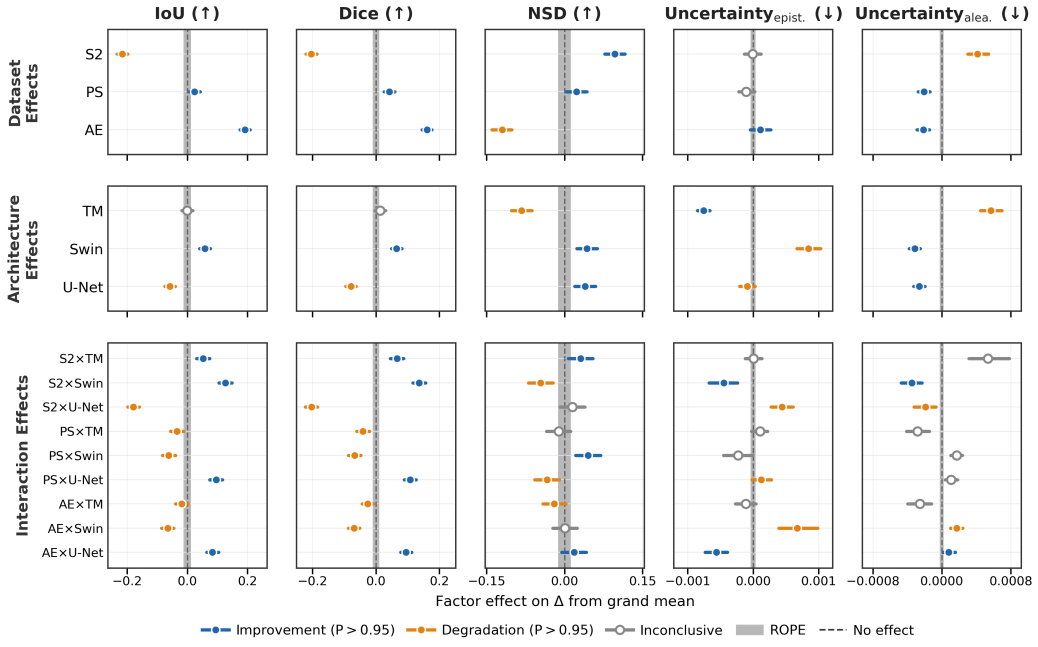

Figure 5. Factorial model posterior estimates for dataset main effects (top), architecture main effects (middle), and dataset-architecture interaction effects (bottom) across five metrics. Points show posterior means; horizontal lines show 95 % HDIs. Colour indicates classification relative to the ROPE: blue indicates improvement vs grand mean ($P > 0.95$), orange indicates degradation ($P > 0.95$), grey indicates an inconclusive effect. The shaded region shows the ROPE. Dashed vertical line at zero (no effect). X-axis: posterior factor effect Δ from grand mean.

NSD, sensor effects are fully inverted, with Sentinel-2 showing a strong positive effect and aerial imagery a strong negative effect, consistent with the sensor-relative NSD tolerance described in Sect. 4.2. The use of TerraMind also results in an elevated aleatoric uncertainty.

Architecture main effects on IoU and Dice are substantially smaller than sensor effects. Swin-U-Net shows a small positive effect (+0.06 IoU; $\text{HDI}_{95\,\%}$ [+0.05, +0.07]) and U-Net a small negative effect (-0.06 IoU; $\text{HDI}_{95\,\%}$ [-0.06, -0.05]), while TerraMind is practically indistinguishable from the grand mean with a degradation of -0.001 IoU, but a probability of falling within the ROPE of 83 %. For NSD, TerraMind shows a strong negative architecture main effect (-0.08; $\text{HDI}_{95\,\%}$ [-0.09, -0.07]), indicating systematically poorer boundary delineation regardless of sensor. For epistemic uncertainty, architecture is the dominant structural driver. TerraMind shows near-zero epistemic uncertainty across all sensors, while Swin-U-Net shows consistently higher parameter uncertainty than U-Net.

Interaction effects show that architecture rankings are sensor-dependent. On aerial and PlanetScope imagery, all interaction terms for IoU and Dice fall within or near the ROPE. The largest interactions occur on Sentinel-2, where the Swin-U-Net

https://doi.org/10.5194/egusphere-2026-4074
Preprint. Discussion started: 20 August 2026

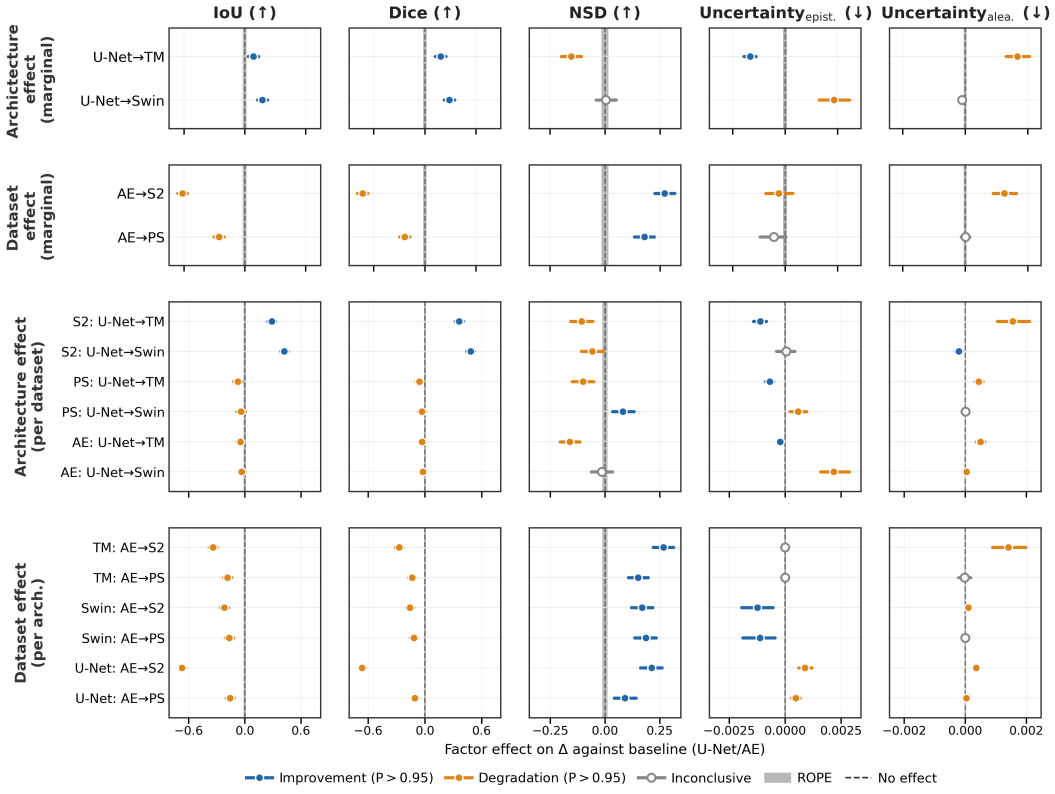

Figure 6. Posterior mean contrasts Δ relative to reference baselines (U-Net for architecture contrasts; AE for dataset contrasts) across five metrics, with 95 % HDIs. Rows show: marginal architecture contrasts averaged over all datasets, marginal dataset contrasts averaged over all architectures, architecture contrasts within each dataset, dataset contrasts within each architecture. Arrow notation (e.g. U-Net $\rightarrow$ TM) denotes the direction of the contrast. Colour indicates posterior classification: blue = improvement ($P > 0.95$), orange = degradation ($P > 0.95$), grey = inconclusive. The shaded region shows the ROPE; dashed line at zero. X-axis: posterior mean contrast Δ from baseline (U-Net / AE).

interaction is strongly positive (+0.13 IoU; $\mathrm{HDI}_{95\,\%}$ [+0.11, +0.14]) and the U-Net interaction strongly negative (-0.18 IoU; $\mathrm{HDI}_{95\,\%}$ [-0.19, -0.17]).

5.3 Sensitivity of architecture choice to spatial resolution

Figure 6 shows posterior contrasts relative to the conventional baseline of U-Net paired with aerial imagery, at both marginal and within-sensor levels. The contrasts indicate that the importance of architecture choice scales with decreasing spatial resolution, while a change in spatial resolution carries a large and architecture-independent cost.

At aerial and PlanetScope resolution, architecture choice is of limited practical consequence. Both Swin-U-Net and TerraMind show small credible degradations relative to U-Net on these sensors for IoU and Dice, while the aerial-to-PlanetScope

https://doi.org/10.5194/egusphere-2026-4074
Preprint. Discussion started: 20 August 2026

IoU reduction is consistent across all three architectures at approximately -0.16 to -0.18 IoU. No architecture therefore offers a meaningful advantage over U-Net when high-quality imagery is available.

The picture changes substantially when handling Sentinel-2 imagery. Here U-Net degrades by -0.67 IoU through a change from aerial to Sentinel-2, compared with -0.22 for Swin-U-Net and -0.34 for TerraMind. This indicates that architecture selection is effectively irrelevant when high-quality imagery is available, but becomes a first-order concern when analyses are constrained to coarse imagery. The within-sensor Swin-U-Net to U-Net contrast on Sentinel-2 alone reaches $+0.42$ IoU, comparable in magnitude to the entire inter-sensor spread observed for any single architecture.

The practical implications of these findings are illustrated in Fig. 7 for a large, but partially overgrown coastal driftwood deposit. On aerial images, all three architectures recover the deposit geometry with high ensemble agreement, though TerraMind shows somewhat lower confidence and misses portions of the deposit boundary, consistent with its marginally lower posterior mean IoU on aerial imagery. On PlanetScope, Swin-U-Net and TerraMind recover the deposit with high confidence, while U-Net struggles to identify it across runs. For Sentinel-2, U-Net produces fragmented low-confidence predictions scattered across the deposit, consistent with mixed-pixel ambiguity at $10\,\mathrm{m}$ resolution between deposit and overgrowth. Swin-U-Net and TerraMind recover the deposit geometry at substantially higher confidence despite the coarse resolution, though both over-predict beyond the deposit boundary, reflecting the difficulty of precise delineation at $10\,\mathrm{m}$.

5.4 Reliability of model selection under single run evaluation

The analyses in the previous sections are based on 10 independently initialised training runs per configuration and a hierarchical model that explicitly accounts for run-to-run variability. Yet, in conventional studies, the reporting of single-point estimates is much more common. To assess what a conventional single-run evaluation would have concluded instead, we compared rank orderings derived from individual training runs against the posterior expected ranks from the hierarchical model, for each individual metric and for a weighted composite score combining all five metrics. Results are shown in Supplementary Fig. S8.

The posterior composite ranking places U-Net paired with aerial imagery first ($\Pr(\mathrm{rank}=1)=0.68$) and Swin-U-Net paired with PlanetScope as the only credible alternative for the top position ($\Pr(\mathrm{rank}=1)=0.28$). The posterior probability that the performance difference between these two configurations falls within the composite equivalence threshold is 0.94, indicating that Swin-U-Net paired with PlanetScope is a practically competitive alternative to the aerial reference in operational contexts where aerial imagery is unavailable. The bottom of the ranking is occupied by TerraMind configurations and U-Net on Sentinel-2, with near-deterministic rank assignments. Full composite ranking probabilities are given in Table 2.

Single-run evaluation recovers these rankings unreliably. For IoU and Dice, exact rank agreement with the posterior ordering averages 38.9 % and 41.1 % respectively ($\tau=0.48$ and 0.47), with instability concentrated in the middle ranks. For comparison, random ordering of ranks would correspond to a rank agreement of 11 %. Most consequentially, the configuration the posterior places at rank 2, Swin-U-Net paired with PlanetScope, is assigned to ranks 4 or 5 in the majority of single IoU and Dice runs. A practitioner relying on single-run evaluation would therefore overlook this combination as a practical alternative entirely. NSD single-run rankings are near-random ($\tau=-0.04$), consistent with the sensitivity of this metric to sensor-relative tolerance documented in Sect. 4.2. Epistemic uncertainty rankings are near-random ($\tau=-0.13$), driven by TerraMind's structurally near-

https://doi.org/10.5194/egusphere-2026-4074
Preprint. Discussion started: 20 August 2026

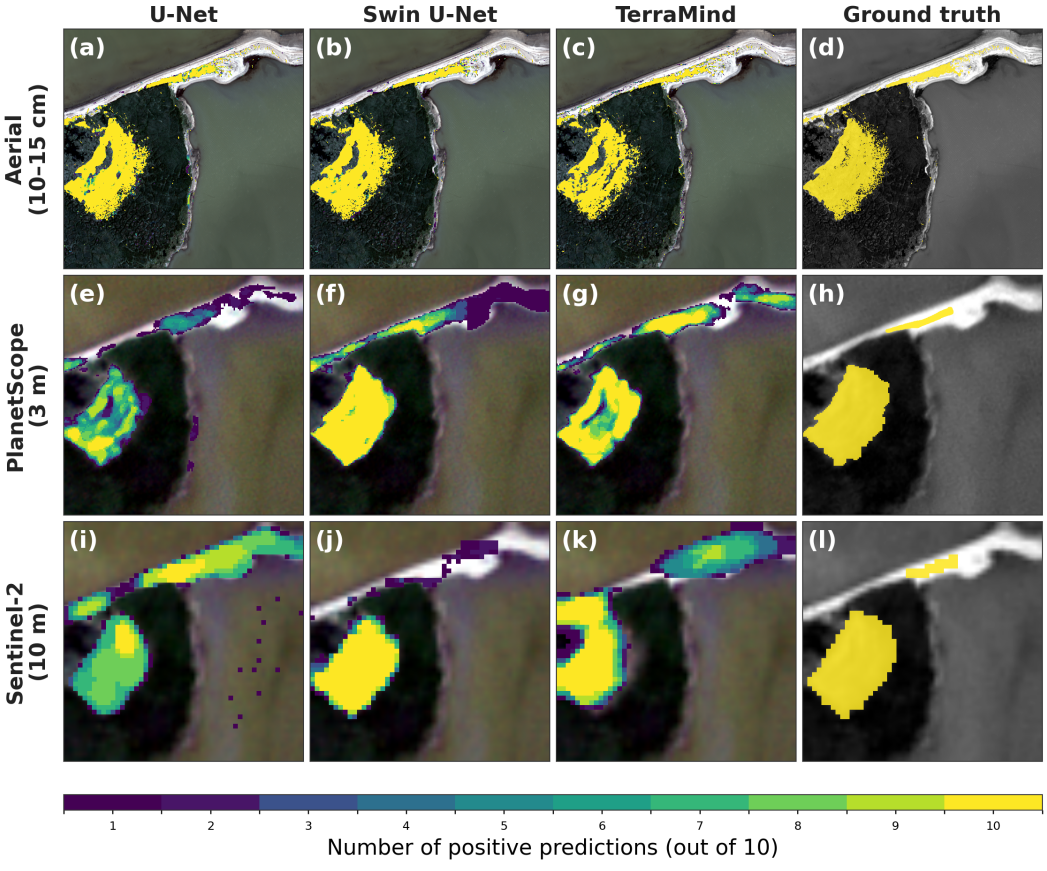

Figure 7. Qualitative comparison of driftwood predictions for a representative test frame across all sensor–architecture combinations. Columns show predictions from U-Net (a, e, i), Swin-U-Net (b, f, j), and TerraMind (c, g, k), alongside the ground-truth label (d, h, l). Rows correspond to aerial imagery at 10–15 cm resolution (a–d), PlanetScope at 3 m (e–h), and Sentinel-2 at 10 m (i–l). The colourmap encodes ensemble agreement across 10 independent training runs; pixels predicted by all 10 runs (yellow) indicate high-confidence detections, while pixels predicted by only 1–3 runs (purple) indicate uncertain or inconsistent detections. Background imagery is displayed in true colour. The ground truth column (d, h, l) shows the binary driftwood label overlaid on a greyscale background.

zero epistemic uncertainty causing it to rank first in all single runs despite its moderate composite rank. Aleatoric uncertainty is the most stable individual metric ($\tau = 0.71$).

Multi-metric aggregation partially stabilises rank estimates. The composite score achieves the highest rank correlation of any individual metric ($\tau = 0.64$) without requiring additional training runs. However, even the composite score falls well short of reliable recovery of the posterior ordering, reinforcing that multi-run evaluation within a probabilistic framework is necessary for robust model selection in this setting.

https://doi.org/10.5194/egusphere-2026-4074
Preprint. Discussion started: 20 August 2026

Table 2. Posterior ranking probabilities, expected rank, and 95 % highest density interval (HDI) for rank, across all nine sensor–architecture combinations, ordered by expected rank.

Configuration	$\Pr(\text{rank} = 1)$	$\Pr(\text{top}\,2)$	$E[\text{rank}]$	rank HDI 95 %
U-Net | AE	0.6844	0.9851	1.331	[1.0, 2.0]
Swin-U-Net | PS	0.2771	0.8698	1.859	[1.0, 3.0]
U-Net | PS	0.0383	0.1409	2.929	[2.0, 4.0]
Swin-U-Net | S2	0.0003	0.0043	3.882	[3.0, 4.0]
Swin-U-Net | AE	0.0000	0.0000	5.089	[5.0, 6.0]
TerraMind | PS	0.0000	0.0000	5.912	[5.0, 6.0]
TerraMind | AE	0.0000	0.0000	7.000	[7.0, 7.0]
TerraMind | S2	0.0000	0.0000	8.000	[8.0, 8.0]
U-Net | S2	0.0000	0.0000	8.999	[9.0, 9.0]

6 Discussion

6.1 Information content determines segmentation performance

The four-fold difference in magnitude between sensor main effects and architecture main effects reflects a fundamental constraint on model performance. No architecture can recover discriminative information lost through spatial averaging. Driftwood deposits are detected primarily through textural rather than spectral information, due to the spectral similarity of bleached driftwood and background in both the visible and near-infrared (Buscombe et al., 2024; Sendrowski and Wohl, 2021). At sub-metre resolution, the spatial structure of individual logs and inter-log gaps provides textural information that any sufficiently expressive model can exploit. As described by Sendrowski and Wohl (2021), wood classification accuracy at very high spatial resolution is governed primarily by boundary clarity rather than spectral separability. With increasing ground sampling distance, the textural signal is progressively lost to spatial averaging, and at 10 m most deposits outside the largest mats are represented as predominantly mixed pixels. The observed sensor hierarchy can therefore be viewed as an information-theoretic ceiling rather than a modelling failure, and should be expected to persist across model families and training strategies.

Within the remote sensing literature, direct IoU comparisons for driftwood mapping are limited. Only Stadie et al. (2025) reports an IoU of 0.578 for a U-Net on PlanetScope, yet our study achieves a posterior mean IoU of 0.641. This gain is largely attributable to the comprehensive hyperparameter search.

One important caveat applies to the interpretation of sensor effects across all comparisons reported here. Because annotations were generated independently at each sensor's native resolution, the observed performance differences absorb both resolution-driven detectability and annotation-scale effects. A deposit labelled as a single polygon at 10 m resolution may correspond to multiple individually resolved objects in the aerial annotation. The sensor effect should therefore be understood as a measure of expected end-to-end operational performance under realistic annotation conditions, rather than as an isolated consequence of spatial resolution alone. For applications where a common label set across sensors is required, this distinction is methodologically significant and should be accounted for in study design.

https://doi.org/10.5194/egusphere-2026-4074
Preprint. Discussion started: 20 August 2026

6.2 Resolution dependence of architecture effects and implications for operational deployment

For aerial and PlanetScope imagery, architecture choice can be described as of limited practical consequence for overlap performance. The HDIs of all three architectures overlap substantially for IoU and Dice at both sensors, and the aerial-to-PlanetScope performance reduction is consistent across architectures at approximately -0.16 to -0.18 IoU. For practitioners working at these resolutions, sensor quality and training data diversity are the primary determinants of segmentation performance, and the added complexity of transformer-based architectures is not justified on pure performance grounds.

The same, however, cannot be said for Sentinel-2. With U-Net on Sentinel-2 achieving a posterior mean IoU of just 0.127, and with the hyperparameter search having explicitly explored dilation rates and feature map dimensionality for this combination, this effect cannot be attributed to unfavourable parameterisation. It rather reflects a mismatch between the inductive bias of convolutional networks and the characteristics of coarse-resolution imagery. At 10 m, Sentinel-2 pixels are highly mixed, local spectral contrast between wood and background is weak, and deposit boundaries are diffuse. In this regime, local receptive fields are not sufficient to determine pixel identity from neighbourhood information alone, producing the spatially fragmented, low-confidence predictions visible in Fig. 7(i). The observation that U-Net on Sentinel-2 nevertheless recovers large spatially distinct deposits supports that the failure is specific to the mixed-pixel regime rather than a uniform property of the architecture. Self-attention mechanisms as used by Swin-U-Net and TerraMind allow the model to use non-local contextual regularities when local evidence is ambiguous (Dosovitskiy et al., 2021; Liu et al., 2022), which likely explains why both architectures degrade substantially more gradually than U-Net with decreasing resolution.

Beyond driftwood mapping, architecture benchmarks are typically conducted at a fixed resolution, and recommendations derived from one sensor or resolution setting are routinely applied to others without independent re-evaluation. Our results, based on three architectures spanning convolutional, transformer, and foundation model paradigms, indicate that this practice can produce substantially misleading conclusions, as the relative ranking of these three architectures reverses across the resolution spectrum. Whether this pattern holds for a broader set of architectures, including hybrid convolutional-attention designs or architectures specifically optimised for coarse-resolution imagery, remains an open question that our experimental design cannot resolve. For any spectrally ambiguous target where local texture is the primary discriminative feature, architecture selection should always be conducted at the resolution of the imagery used in operational deployment rather than inherited from benchmarks at a different resolution (Henderson et al., 2018; Kattenborn et al., 2022).

For practitioners in the driftwood mapping community, these results translate into concrete guidance. For aerial imagery, U-Net remains the best-performing architecture and the added complexity of transformer-based alternatives does not yield any significant improvement. Where aerial campaigns are logistically or financially impractical, Swin-U-Net combined with PlanetScope imagery provides the most competitive accessible alternative, with a posterior mean IoU of 0.600 and the highest composite rank among all satellite-based configurations. This combination benefits from PlanetScope's near-daily global coverage, enabling large-scale and temporally flexible monitoring without dedicated flight campaigns, though this approach will underestimate total driftwood area due to individual logs falling below the detection threshold (Stadie et al., 2025). For practitioners working in depositional environments spectrally and morphologically similar to the Mackenzie Delta and constrained to

https://doi.org/10.5194/egusphere-2026-4074
Preprint. Discussion started: 20 August 2026

Sentinel-2, U-Net is likely to underperform substantially relative to transformer-based alternatives. In this context, Swin-U-Net represents the more robust default choice, with TerraMind as a viable alternative where cross-sensor consistency or reduced training variance is a priority. Whether this ranking holds in driftwood systems with different background characteristics, such as the rocky coastlines of Norway or Iceland, requires independent empirical evaluation.

6.3 Foundation model performance in narrow target applications

TerraMind offers two properties that distinguish it from the task-specific architectures evaluated here. It produces near-zero epistemic uncertainty across all sensor configurations, making predictions highly reproducible across training runs, and its degradation from aerial to Sentinel-2 resolution is substantially smaller than that of U-Net. Both properties are consistent with the design rationale of Earth observation foundation models, where pre-trained representations provide a stable inductive prior that reduces sensitivity to training stochasticity and domain shift (Bommasani et al., 2022; Jakubik et al., 2023).

However, TerraMind does not outperform its competitors on Sentinel-2 despite having been pre-trained on that sensor, and produces substantially elevated aleatoric uncertainty on Sentinel-2 relative to both alternatives. This can be attributed to a domain gap existing between the pre-training distribution and the target task, with TerraMind's pretraining dataset being optimised for globally diverse land cover (Jakubik et al., 2025). Driftwood as described here, being spectrally similar to sand and spatially restricted to Arctic and boreal coastlines, has almost certainly no analogue in that distribution. Therefore, the decoder must bridge this gap from a very limited fine-tuning dataset.

Despite this, however, TerraMind has shown considerable performance across the resolution spectrum. Although it does not outperform its specialised competitors, its performance on high-resolution imagery was nonetheless comparable to those of U-Net and Swin-U-Net alike. Considering that those image types were not part of the initial pretraining of the foundation model, its performance is remarkable. This confirms that foundation models can serve as a potent backbone for difficult segmentation tasks, even outside of their native training domain.

Yet, as pre-trained models become increasingly accessible through platforms such as TerraTorch (Gomes et al., 2025), their adoption risks being driven by general assumptions of superiority rather than by empirical evaluation on the target task. Our results suggest that foundation model pre-training does not guarantee superior performance on spectrally unusual targets that are likely absent from the pre-training distribution, with TerraMind failing to outperform well-tuned task-specific alternatives on any sensor despite Sentinel-2 being its native training domain. This supports the broader argument that foundation models cannot be assumed to automatically extrapolate within a given modality, and that their performance limits in out-of-distribution settings require empirical investigation rather than prior assumption (Bommasani et al., 2022). Foundation models should therefore be evaluated empirically against task-specific alternatives, particularly in narrow-target applications where the fine-tuning dataset may be insufficient to bridge the gap between the pre-training distribution and the target domain.

6.4 Implications for benchmarking practice in environmental remote sensing

The rank instability analysis shows that conventional single-run evaluation fails to recover reliable model rankings in this setting, with exact rank agreement below 42 % for IoU and Dice and near-random agreement for NSD. Due to the spectral

https://doi.org/10.5194/egusphere-2026-4074
Preprint. Discussion started: 20 August 2026

ambiguity of the segmentation target, the sensitivity of model training to random weight initialisation is increased and limited training data amplify the influence of stochastic data sampling (Henderson et al., 2018), while strong class imbalance makes augmentation choices a substantial source of run-to-run variability (Buda et al., 2018).

In practice, this instability can produce qualitative errors in model selection. The posterior analysis identifies Swin-U-Net paired with PlanetScope as a practically competitive alternative to the aerial reference, with a 94 % posterior probability that its composite performance falls within the equivalence threshold of the top-ranked configuration. Single-run IoU and Dice evaluations consistently assign this combination to ranks 4 or 5, which would lead a practitioner to overlook it entirely and conclude that no accessible satellite-based alternative exists.

As shown, NSD rankings are near-random due to sensitivity to the sensor-relative tolerance width, and epistemic uncertainty rankings are near-inverted because TerraMind's structurally near-zero epistemic uncertainty causes it to rank first on that metric in every single run despite its moderate composite rank. While multi-metric aggregation partially stabilises rank estimates, this does not constitute reliable model selection without multi-run evaluation. The Bayesian hierarchical framework used here addresses this by treating run-to-run variability as a quantity to be modelled and propagated into all downstream comparisons and rankings.

For the driftwood mapping community, this has a direct practical consequence, as we found that Swin-U-Net paired with PlanetScope constitutes a competitive alternative to aerial acquisition, which would have been missed entirely under conventional evaluation. We therefore argue that multi-run evaluation within a probabilistic framework should be considered a minimum methodological standard for segmentation benchmarking in environmental remote sensing, particularly in settings characterised by limited training data, spectral ambiguity, and strong class imbalance (Bouthillier et al., 2021; Henderson et al., 2018; Kattenborn et al., 2022; Reinke et al., 2024).

6.5 Limitations and outlook

Several limitations of this study should be acknowledged when interpreting the results and applying the recommendations derived from them. All training and evaluation data originate from a single geographic region which, while representing the world's largest documented driftwood system (Sendrowski et al., 2023), has a specific combination of target properties and confounding backgrounds. The dominant background surfaces in the Mackenzie Delta may differ from those encountered in other major driftwood systems such as, for example, the Norwegian and Icelandic coastlines. While the sensor hierarchy is grounded in information-theoretic constraints that should generalise broadly, the specific magnitudes of architecture effects and their sensor dependence reflect the spectral and morphological characteristics of this region, and validation in geographically distinct contexts is a necessary next step.

A second limitation concerns the temporal mismatch between sensors. Although acquisition offsets were minimised by design, with mean absolute differences of 1.1 days for PlanetScope and 2.4 days for Sentinel-2 relative to the aerial campaign, the Mackenzie Delta is a dynamic system subject to wood transport by tidal forcing, storm surges, and intermittent wood floods with a return interval of approximately five years (Kramer et al., 2017; Sendrowski et al., 2023). Even short temporal windows cannot guarantee identical deposit geometries across sensors, and the resulting label-to-image misalignment may

https://doi.org/10.5194/egusphere-2026-4074
Preprint. Discussion started: 20 August 2026

affect boundary-sensitive metrics. Weather conditions during the acquisition period were calm and stable and no storm surges or wood-flood events were recorded in the region during this window, making substantial deposit displacement within these offsets unlikely. Beyond small temporal offsets, the considered sensors also differ spectrally and radiometrically. Sentinel-2 benefits from stable, well-characterised calibration, while MACS aerial image acquisitions are subject to greater illumination variability (Rettelbach et al., 2024), meaning observed performance gaps reflect not only spatial resolution but also the radiometric properties of each imaging system.

Finally, this study evaluates a purposeful but limited selection of architectures and sensors. The Bayesian multi-run evaluation framework is directly applicable to further architecture comparisons, including hybrid convolutional-attention designs and architectures specifically designed for coarse-resolution or mixed-pixel imagery. Its extension to related Arctic mapping problems involving comparably sparse and spectrally challenging targets such as retrogressive thaw slumps, ice-wedge polygon networks, or thermokarst lake margins represents a promising direction for follow-on work as the community's focus expands towards systematic pan-Arctic land-surface monitoring.

7 Conclusions

We evaluated three deep learning architectures across aerial, PlanetScope, and Sentinel-2 imagery for driftwood segmentation in the Mackenzie Delta using a Bayesian hierarchical framework that explicitly accounts for run-to-run variability in stochastic neural network training.

Sensor modality was the dominant performance driver, outweighing architecture choice by a factor of four across overlap metrics. For driftwood detection, architecture selection was of limited consequence when high-quality imagery was available, but became critical at coarser resolution, where U-Net showed worse performance, while transformer-based architectures degraded more gradually. The TerraMind foundation model did not outperform task-specific architectures on any sensor, including Sentinel-2, its native pretraining modality. Given that all training and evaluation data originate from a single geographic region, the absolute magnitudes of these effects and their sensor dependence should be interpreted as region-specific estimates pending cross-regional validation.

Based on our results we can provide concrete guidance for future work. Aerial imagery with U-Net remains a strong-performing configuration where multispectral aerial imagery is available or acquisition is feasible, while Swin-U-Net paired with PlanetScope is the most competitive accessible alternative for large-scale operational monitoring. For Sentinel-2, transformer-based architectures should be preferred over U-Net regardless of computational cost. These conclusions were reachable only through multi-run probabilistic evaluation. The evaluation framework applied here is directly transferable to other spectrally ambiguous targets in Arctic and environmental remote sensing where training data are limited and reliable model selection is critical.

Code availability. All code used to derive the presented results is deposited on Zenodo: https://doi.org/10.5281/zenodo.20824839.

https://doi.org/10.5194/egusphere-2026-4074
Preprint. Discussion started: 20 August 2026

Data availability. Annotation files for all three sensor modalities are available at https://doi.org/10.5281/zenodo.20825411. Sentinel-2 imagery is freely available from the Copernicus Data Space; scene identifiers and acquisition dates are provided in Supplementary Table S9. Aerial imagery from the Perma-X 2025 campaign will be deposited to PANGAEA upon publication. PlanetScope imagery cannot be redistributed under the terms of the research licence; scene identifiers sufficient for licensed users to reproduce the dataset are provided in Supplementary Table S9.

Author contributions. CS, GG, BD and MB designed the study. IN, GG, and CS conducted the airborne surveys. IN processed the aerial imagery. CS implemented the analysis, analyzed the results and drafted the manuscript, with contributions from all authors. GG, BD and MB supervised the project. All authors reviewed and approved the final manuscript.

Competing interests. The authors declare that they have no conflict of interest.

Statement on inclusion in global research. This study was conducted by European-based researchers in the Canadian Arctic, on Inuvialuit traditional territory. The airborne survey was carried out under permits granted following consultation and approval by Inuvialuit and Gwich'in governance bodies, including Hunters and Trappers Committees, covering environmental impact screening and authorisation within the territory. The dataset, model weights, and code will be openly available at no cost upon publication, and we commit to engaging Inuvialuit and Gwich'in organisations on this work's relevance to land and resource management.

Acknowledgements. AWI base funding was used to conduct the airborne campaign P5_258 with the AWI Polar-5 airplane. We acknowledge support for the airborne surveys in this study by AWI Logistics and Scientific Platforms personnel, specifically Daniel Steinhage, Benjamin Harting, Christoph Petersen, Martin Gehrmann, Maximilian Stöhr, Cristina Sans Coll, Eduard Gebhard, Silke Henkel, and Uwe Nixdorf, who facilitated the Polar 5 airplane logistics and technical preparation. We thank the pilot crew (Dean Emberley, Almina Kovcic, and Dwayne Bailey) of Kenn Borek Air for their excellent service during the survey flights based out of Inuvik Airport. We thank DLR for designing and providing the MACS. This work was directly supported by a HEIBRiDS grant to CS. GG and IN were partially supported by ML4Earth, Google PDG, Helmholtz Foundation Model Initiative project 3D-ABC, and the PeTCaT project funded by Schmidt Sciences.

Generative AI tools were used to assist with this manuscript. GitHub Copilot (GitHub, Microsoft) was used to assist with code commenting and refinement during the development of analysis and processing scripts. Claude (Anthropic) was used to assist with language improvement and correction at the initial stage of the manuscript text. The final version of the code and manuscript has been subsequently reviewed, revised, and approved by all authors.

https://doi.org/10.5194/egusphere-2026-4074
Preprint. Discussion started: 20 August 2026

References

Agarap, A. F.: Deep Learning using Rectified Linear Units (ReLU), 2018.

Alix, C.: Deciphering the impact of change on the driftwood cycle: contribution to the study of human use of wood in the Arctic, Glob. Planet. Change, 47, 83–98, https://doi.org/10.1016/j.gloplacha.2004.10.004, 2005.

Benavoli, A., Corani, G., Demšar, J., and Zaffalon, M.: Time for a Change: A Tutorial for Comparing Multiple Classifiers Through Bayesian Analysis, J. Mach. Learn. Res., 18, 1–36, 2017.

Bergstra, J., Bardenet, R., Bengio, Y., and Kégl, B.: Algorithms for Hyper-Parameter Optimization, in: Advances in Neural Information Processing Systems, pp. 2546–2554, https://doi.org/10.5555/2986459.2986743, 2011.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., Arx, S. v., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X. L., Li, X., Ma, T., Malik, A., Manning, C. D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J. C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J. S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A. W., Tramèr, F., Wang, R. E., et al.: On the Opportunities and Risks of Foundation Models, arXiv [preprint], https://doi.org/10.48550/arXiv.2108.07258, 2022.

Bouthillier, X., Delaunay, P., Bronzi, M., Trofimov, A., Nichyporuk, B., Szeto, J., Mohammadi Sepahvand, N., Raff, E., Madan, K., Voleti, V., Ebrahimi Kahou, S., Michalski, V., Arbel, T., Pal, C., Varoquaux, G., and Vincent, P.: Accounting for Variance in Machine Learning Benchmarks, in: Proceedings of Machine Learning and Systems, pp. 747–769, 2021.

Buda, M., Maki, A., and Mazurowski, M. A.: A systematic study of the class imbalance problem in convolutional neural networks, Neural Netw., 106, 249–259, https://doi.org/10.1016/j.neunet.2018.07.011, 2018.

Büntgen, U., Oppenheimer, C., Trnka, M., Kempf, M., Holman, I., Arosio, T., Bebchuk, T., and Esper, J.: Arctic driftwood proposal for durable carbon removal, Npj Clim. Action, 5, 1, https://doi.org/10.1038/s44168-025-00327-1, 2026.

Burn, C. R. and Kokelj, S. V.: The environment and permafrost of the Mackenzie Delta area, Permafr. Periglac. Process., 20, 83–105, https://doi.org/10.1002/ppp.655, 2009.

Buscombe, D., Warrick, J., Ritchie, A., East, A., McHenry, M., McCoy, R., Foxgrover, A., and Wohl, E.: Remote Sensing Large-Wood Storage Downstream of Reservoirs During and After Dam Removal: Elwha River, Washington, USA, Earth Space Sci., 11, https://doi.org/10.1029/2024ea003544, 2024.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., and Wang, M.: Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation, in: Computer Vision – ECCV 2022 Workshops, pp. 205–218, Springer, Cham, https://doi.org/10.1007/978-3-031-25066-8_9, 2021.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H.: Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation, arXiv [preprint], https://doi.org/10.48550/ARXIV.1802.02611, 2018.

Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., Burke, M., Lobell, D. B., and Ermon, S.: SatMAE: Pre-training Transformers for Temporal and Multi-Spectral Satellite Imagery, arXiv [preprint], https://doi.org/10.48550/ARXIV.2207.08051, 2022.

https://doi.org/10.5194/egusphere-2026-4074
Preprint. Discussion started: 20 August 2026

Corani, G., Benavoli, A., Demšar, J., Mangili, F., and Zaffalon, M.: Statistical comparison of classifiers through Bayesian hierarchical modelling, Mach. Learn., 106, 1817–1837, https://doi.org/10.1007/s10994-017-5641-9, 2017.

Dalaiden, Q., Goosse, H., Lecomte, O., and Docquier, D.: A model to interpret driftwood transport in the Arctic, Quat. Sci. Rev., 191, 89–100, https://doi.org/10.1016/j.quascirev.2018.05.004, 2018.

Davidson, S. G., Hesp, P. A., and Silva, G. M. D.: Controls on dune scarping, Prog. Phys. Geogr. Earth Environ., 44, 923–947, https://doi.org/10.1177/0309133320932880, 2020.

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

Dice, L. R.: Measures of the Amount of Ecologic Association Between Species, Ecology, 26, 297–302, https://doi.org/10.2307/1932409, 1945.

Dietenberger, S., Mueller, M., Stöcker, B., Dubois, C., Arlaud, H., Adam, M., Hese, S., Meyer, H., and Thiel, C.: Accurate Mapping of Downed Deadwood in a Dense Deciduous Forest Using UAV-SfM Data and Deep Learning, Remote Sens., https://doi.org/10.3390/rs17091610, 2025.

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, arXiv [preprint], https://doi.org/10.48550/arXiv.2010.11929, 2021.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F., and Bargellini, P.: Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services, Remote Sens. Environ., 120, 25–36, https://doi.org/10.1016/j.rse.2011.11.026, 2012.

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., and Gelman, A.: Visualization in Bayesian workflow, J. R. Stat. Soc. Ser. A Stat. Soc., 182, 389–402, https://doi.org/10.1111/rssa.12378, 2019.

Gal, Y. and Ghahramani, Z.: Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning, in: Proceedings of The 33rd International Conference on Machine Learning, pp. 1050–1059, 2016.

Gelman, A.: Analysis of variance—why it is more important than ever, Ann. Stat., 33, 1–53, https://doi.org/10.1214/009053604000001048, 2005.

Gelman, A.: Prior distributions for variance parameters in hierarchical models, Bayesian Anal., 1, 515–533, https://doi.org/10.1214/06-BA117A, 2006.

Gomes, C., Blumenstiel, B., De Sousa Almeida, J. L., De Oliveira, P. H., Fraccaro, P., Escofet, F. M., Szwarcman, D., Simumba, N., Kienzler, R., and Zadrozny, B.: TerraTorch: The Geospatial Foundation Models Toolkit, in: IGARSS 2025 - 2025 IEEE International Geoscience and Remote Sensing Symposium, pp. 6364–6368, https://doi.org/10.1109/IGARSS55030.2025.11243570, 2025.

Grilliot, M. J., Walker, I. J., and Bauer, B. O.: Aeolian sand transport and deposition patterns within a large woody debris matrix fronting a foredune, Geomorphology, 338, 1–15, https://doi.org/10.1016/j.geomorph.2019.04.010, 2019.

Grosse, G., Nitze, I., Stadie, C., and von Baeckmann, C.: Master tracks in different resolutions during POLAR 5 campaign P5-258_PERMA-X_2025, PANGAEA [data set], https://doi.org/10.1594/PANGAEA.986494, 2025.

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., and Meger, D.: Deep Reinforcement Learning That Matters, Proc. AAAI Conf. Artif. Intell., 32, https://doi.org/10.1609/aaai.v32i1.11694, 2018.

Hoffman, M. D. and Gelman, A.: The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo, J. Mach. Learn. Res., 15, 1593–1623, 2014.

https://doi.org/10.5194/egusphere-2026-4074
Preprint. Discussion started: 20 August 2026

Husson, E., Ecke, F., and Reese, H.: Comparison of Manual Mapping and Automated Object-Based Image Analysis of Non-Submerged Aquatic Vegetation from Very-High-Resolution UAS Images, Remote Sens., 8, 724, https://doi.org/10.3390/rs8090724, 2016.

Jaccard, P.: The Distribution of the Flora in the Alpine Zone, New Phytol., 11, 37–50, https://doi.org/10.1111/j.1469-8137.1912.tb05611.x, 1912.

Jakubik, J., Roy, S., Phillips, C. E., Fraccaro, P., Godwin, D., Zadrozny, B., Szwarcman, D., Gomes, C., Nyirjesy, G., Edwards, B., Kimura, D., Simumba, N., Chu, L., Mukkavilli, S. K., Lambhate, D., Das, K., Bangalore, R., Oliveira, D., Muszynski, M., Ankur, K., Ramasubramanian, M., Gurung, I., Khallaghi, S., Li, H., Cecil, M., Ahmadi, M., Kordi, F., Alemohammad, H., Maskey, M., Ganti, R., Weldemariam, K., and Ramachandran, R.: Foundation Models for Generalist Geospatial Artificial Intelligence, arXiv [preprint], https://doi.org/10.48550/arXiv.2310.18660, 2023.

Jakubik, J., Yang, F., Blumenstiel, B., Scheurer, E., Sedona, R., Maurogiovanni, S., Bosmans, J., Dionelis, N., Marsocci, V., Kopp, N., Ramachandran, R., Fraccaro, P., Brunschwiler, T., Cavallaro, G., Bernabé-Moreno, J., and Longépé, N.: TerraMind: Large-Scale Generative Multimodality for Earth Observation, arXiv [preprint], https://doi.org/10.48550/arxiv.2504.11171, 2025.

Kattenborn, T., Schiefer, F., Frey, J., Feilhauer, H., Mahecha, M. D., and Dormann, C. F.: Spatially autocorrelated training and validation samples inflate performance assessment of convolutional neural networks, ISPRS Open J. Photogramm. Remote Sens., 5, 100 018, https://doi.org/10.1016/j.ophoto.2022.100018, 2022.

Kendall, A. and Gal, Y.: What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?, in: Advances in Neural Information Processing Systems, pp. 5574–5584, https://doi.org/10.5555/3295222.3295309, 2017.

Kohavi, R.: A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection, in: Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI), pp. 1137–1145, 1995.

Kramer, N., Wohl, E., Hess-Homeier, B., and Leisz, S.: The pulse of driftwood export from a very large forested river basin over multiple time scales, Slave River, Canada, Water Resour. Res., 53, 1928–1947, https://doi.org/10.1002/2016WR019260, 2017.

Kruschke, J. K.: Rejecting or Accepting Parameter Values in Bayesian Estimation, Adv. Methods Pract. Psychol. Sci., 1, 270–280, https://doi.org/10.1177/2515245918771304, 2018.

Lehmann, F., Berger, R., Brauchle, J., Hein, D., Meissner, H., Pless, S., Strackenbrock, B., and Wieden, A.: MACS - Modulares Luftbildkamera-System für die Erzeugung hochauflösender photogrammetrischer Produkte, Photogramm. - Fernerkund. - Geoinformation, 2011, 435–446, https://doi.org/10.1127/1432-8364/2011/0096, 2011.

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., and Talwalkar, A.: Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization, J. Mach. Learn. Res., 18, 1–52, 2018.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B.: Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, 2021 IEEE/CVF Int. Conf. Comput. Vis. (ICCV), pp. 9992–10 002, https://doi.org/10.1109/iccv48922.2021.00986, 2022.

Lu, S., Guo, J., Zimmer-Dauphinee, J. R., Nieusma, J. M., Wang, X., VanValkenburgh, P., Wernke, S. A., and Huo, Y.: Vision Foundation Models in Remote Sensing: A Survey, arXiv [preprint], https://doi.org/10.48550/arXiv.2408.03464, 2025.

Mañas, O., Lacoste, A., Giro-i Nieto, X., Vazquez, D., and Rodriguez, P.: Seasonal Contrast: Unsupervised Pre-Training from Uncurated Remote Sensing Data, arXiv [preprint], https://doi.org/10.48550/ARXIV.2103.16607, 2021.

MacLeod, R. F. and Dallimore, S. R.: Assessment of Storm Surge History as Recorded by Driftwood in the Mackenzie Delta and Tuktoyaktuk Coastlands, Arctic Canada, Front. Earth Sci., 9, 698 660, https://doi.org/10.3389/feart.2021.698660, 2021.

Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., Arbel, T., Bogunovic, H., Bradley, A. P., Carass, A., Feldmann, C., Frangi, A. F., Full, P. M., Van Ginneken, B., Hanbury, A., Honauer, K., Kozubek, M., Landman, B. A., März, K., Maier, O.,

https://doi.org/10.5194/egusphere-2026-4074
Preprint. Discussion started: 20 August 2026

Maier-Hein, K., Menze, B. H., Müller, H., Neher, P. F., Niessen, W., Rajpoot, N., Sharp, G. C., Sirinukunwattana, K., Speidel, S., Stock, C., Stoyanov, D., Taha, A. A., Van Der Sommen, F., Wang, C.-W., Weber, M.-A., Zheng, G., Jannin, P., and Kopp-Schneider, A.: Why rankings of biomedical image analysis competitions should be interpreted with care, Nat. Commun., 9, 5217, https://doi.org/10.1038/s41467-018-07619-7, 2018.

Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M. D., Buettner, F., Christodoulou, E., Glocker, B., Isensee, F., Kleesiek, J., Kozubek, M., Reyes, M., Riegler, M. A., Wiesenfarth, M., Kavur, A. E., Sudre, C. H., Baumgartner, M., Eisenmann, M., Heckmann-Nötzel, D., Rädsch, T., Acion, L., Antonelli, M., Arbel, T., Bakas, S., Benis, A., Blaschko, M. B., Cardoso, M. J., Cheplygina, V., Cimini, B. A., Collins, G. S., Farahani, K., Ferrer, L., Galdran, A., Van Ginneken, B., Haase, R., Hashimoto, D. A., Hoffman, M. M., Huisman, M., Jannin, P., Kahn, C. E., Kainmueller, D., Kainz, B., Karargyris, A., Karthikesalingam, A., Kofler, F., Kopp-Schneider, A., Kreshuk, A., Kurc, T., Landman, B. A., Litjens, G., Madani, A., Maier-Hein, K., Martel, A. L., Mattson, P., Meijering, E., Menze, B., Moons, K. G. M., Müller, H., Nichyporuk, B., Nickel, F., Petersen, J., Rajpoot, N., Rieke, N., Saez-Rodriguez, J., Sánchez, C. I., Shetty, S., Van Smeden, M., Summers, R. M., Taha, A. A., Tiulpin, A., Tsaftaris, S. A., Van Calster, B., Varoquaux, G., and Jäger, P. F.: Metrics reloaded: recommendations for image analysis validation, Nat. Methods, 21, 195–212, https://doi.org/10.1038/s41592-023-02151-z, 2024.

Murphy, E., Cornett, A., Nistor, I., and S, B.: Modelling fate and transport of woody debris in coastal waters, in: Int. Conf. Coastal Eng., vol. 10, 2020.

Murphy, E., Nistor, I., Cornett, A., Wilson, J., and Pilechi, A.: Fate and transport of coastal driftwood: A critical review, Mar. Pollut. Bull., 170, 112 649, https://doi.org/10.1016/j.marpolbul.2021.112649, 2021.

Nikolov, S., Blackwell, S., Mendes, R., De Fauw, J., Meyer, C., Hughes, C., Askham, H., Romera-Paredes, B., Karthikesalingam, A., Chu, C., et al.: Clinically Applicable Segmentation of Head and Neck Anatomy for Radiotherapy: Deep Learning Algorithm Development and Validation Study, J. Med. Internet Res., 23, e26 151, https://doi.org/10.2196/26151, 2021.

Ostmeier, S., Axelrod, B., Bertels, J., Isensee, F., Lansberg, M. G., Christensen, S., Albers, G. W., Li, L.-J., and Heit, J. J.: USE-Evaluator: Performance metrics for medical image segmentation models supervised by uncertain, small or empty reference annotations in neuroimaging, Med. Image Anal., 90, 102 927, https://doi.org/10.1016/j.media.2023.102927, 2023.

Pala, A., Oleynik, A., Utseth, I., and Handegard, N. O.: Addressing class imbalance in deep learning for acoustic target classification, ICES J. Mar. Sci., 80, 2530–2544, https://doi.org/10.1093/icesjms/fsad165, 2023.

Pawłuszek-Filipiak, K. and Borkowski, A.: On the Importance of Train–Test Split Ratio of Datasets in Automatic Landslide Detection by Supervised Classification, Remote Sens., 12, 3054, https://doi.org/10.3390/rs12183054, 2020.

Piliouras, A. and Rowland, J. C.: Arctic River Delta Morphologic Variability and Implications for Riverine Fluxes to the Coast, J. Geophys. Res. Earth Surf., 125, e2019JF005 250, https://doi.org/10.1029/2019JF005250, 2020.

Planet Labs PBC: Planet Imagery Product Specifications, 2023.

Reimers, N. and Gurevych, I.: Why Comparing Single Performance Scores Does Not Allow to Draw Conclusions About Machine Learning Approaches, arXiv [preprint], https://doi.org/10.48550/ARXIV.1803.09578, 2018.

Reinke, A., Tizabi, M. D., Baumgartner, M., Eisenmann, M., Heckmann-Nötzel, D., Kavur, A. E., Rädsch, T., Sudre, C. H., Acion, L., Antonelli, M., Arbel, T., Bakas, S., Benis, A., Buettner, F., Cardoso, M. J., Cheplygina, V., Chen, J., Christodoulou, E., Cimini, B. A., Farahani, K., Ferrer, L., Galdran, A., Van Ginneken, B., Glocker, B., Godau, P., Hashimoto, D. A., Hoffman, M. M., Huisman, M., Isensee, F., Jannin, P., Kahn, C. E., Kainmueller, D., Kainz, B., Karargyris, A., Kleesiek, J., Kofler, F., Kooi, T., Kopp-Schneider, A., Kozubek, M., Kreshuk, A., Kurc, T., Landman, B. A., Litjens, G., Madani, A., Maier-Hein, K., Martel, A. L., Meijering, E., Menze, B., Moons, K. G. M., Müller, H., Nichyporuk, B., Nickel, F., Petersen, J., Rafelski, S. M., Rajpoot, N., Reyes, M., Riegler, M. A., Rieke, N., Saez-

https://doi.org/10.5194/egusphere-2026-4074
Preprint. Discussion started: 20 August 2026

Rodriguez, J., Sánchez, C. I., Shetty, S., Summers, R. M., Taha, A. A., Tiulpin, A., Tsaftaris, S. A., Van Calster, B., Varoquaux, G., Yaniv, Z. R., Jäger, P. F., and Maier-Hein, L.: Understanding metric-related pitfalls in image analysis validation, Nat. Methods, 21, 182–194, https://doi.org/10.1038/s41592-023-02150-0, 2024.

Rettelbach, T., Nitze, I., Grünberg, I., Hammar, J., Schäffler, S., Hein, D., Gessner, M., Bucher, T., Brauchle, J., Hartmann, J., Sachs, T., Boike, J., and Grosse, G.: Very high resolution aerial image orthomosaics, point clouds, and elevation datasets of select permafrost landscapes in Alaska and northwestern Canada, Earth Syst. Sci. Data, https://doi.org/10.5194/essd-16-5767-2024, 2024.

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, pp. 234–241, 2015.

Ruiz Villanueva, V., Aarnink, J., Ghaffarian, H., Gibaja del Hoyo, J., Finch, B., Hortobágyi, B., Vuaridel, M., and Piégay, H.: Current progress in quantifying and monitoring instream large wood supply and transfer in rivers, Earth Surf. Process. Landf., 49, 256–276, https://doi.org/10.1002/esp.5765, 2024.

Saisana, M., Saltelli, A., and Tarantola, S.: Uncertainty and sensitivity analysis techniques as tools for the quality assessment of composite indicators, J. R. Stat. Soc. Ser. A Stat. Soc., 168, 307–323, https://doi.org/10.1111/j.1467-985X.2005.00350.x, 2005.

Salehi, S. S. M., Erdogmus, D., and Gholipour, A.: Tversky Loss Function for Image Segmentation Using 3D Fully Convolutional Deep Networks, in: Machine Learning in Medical Imaging, edited by Wang, Q., Shi, Y., Suk, H.-I., and Suzuki, K., vol. 10541, pp. 379–387, Springer, Cham, https://doi.org/10.1007/978-3-319-67389-9_44, 2017.

Schad, D. J., Vasishth, S., Hohenstein, S., and Kliegl, R.: How to capitalize on a priori contrasts in linear (mixed) models: A tutorial, J. Mem. Lang., 110, 104 038, https://doi.org/10.1016/j.jml.2019.104038, 2020.

Sendrowski, A. and Wohl, E.: Remote sensing of large wood in high-resolution satellite imagery: Design of an automated classification work-flow for multiple wood deposit types, Earth Surf. Process. Landf., 46, 2333–2348, https://doi.org/10.1002/esp.5179, 2021.

Sendrowski, A., Wohl, E., Hilton, R., Kramer, N., and Ascough, P.: Wood-Based Carbon Storage in the Mackenzie River Delta: The World's Largest Mapped Riverine Wood Deposit, Geophys. Res. Lett., 50, e2022GL100 913, https://doi.org/10.1029/2022GL100913, 2023.

Snoek, J., Larochelle, H., and Adams, R. P.: Practical Bayesian optimization of machine learning algorithms, in: Proceedings of the 26th International Conference on Neural Information Processing Systems – Volume 2, pp. 2951–2959, Lake Tahoe, Nevada, 2012.

Stadie, C., Brandt, M., Nitze, I., Tong, X., Liu, S., Kariryaa, A., Li, S., Reiner, F., Rettelbach, T., and Grosse, G.: Large driftwood accumulations along arctic coastlines and rivers, Sci. Rep., 15, https://doi.org/10.1038/s41598-025-17426-y, 2025.

Wesche, C., Steinhage, D., and Nixdorf, U.: Polar aircraft Polar5 and Polar6 operated by the Alfred Wegener Institute, J. Large-Scale Res. Facil. JLSRF, 2, A87, https://doi.org/10.17815/jlsrf-2-153, 2016.

Wohl, E., Maximenko, N., and Helm, R.: Global wood cascades from terrestrial sources to terrestrial, freshwater, and marine sinks, Earth-Sci. Rev., 276, 105 425, https://doi.org/10.1016/j.earscirev.2026.105425, 2026.

Woo, M.-K. and Thorne, R.: Streamflow in the Mackenzie Basin, Canada, ARCTIC, 56, 328–340, https://doi.org/10.14430/arctic630, 2003.

Xiao, T., Liu, Y., Zhou, B., Jiang, Y., and Sun, J.: Unified Perceptual Parsing for Scene Understanding, in: Computer Vision – ECCV 2018, edited by Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y., vol. 11209, pp. 432–448, Springer International Publishing, Cham, https://doi.org/10.1007/978-3-030-01228-1_26, 2018.

Zhao, Z., Fan, C., and Liu, L.: GeoSAM: A QGIS plugin using Segment Anything Model (SAM) to accelerate geospatial image segmentation, Zenodo [code], https://doi.org/10.5281/ZENODO.8191039, 2023.

Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., and Fraundorfer, F.: Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources, IEEE Geosci. Remote Sens. Mag., 5, 8–36, https://doi.org/10.1109/MGRS.2017.2762307, 2017.