



SPUN: Deep learning for continuous snow cover fraction retrieval in marginal environments

Leam Howe ¹, Richard Essery ¹, and Elliot J. Crowley ²

¹School of Geosciences, University of Edinburgh, Edinburgh, UK

²School of Engineering, University of Edinburgh, Edinburgh, UK

Correspondence: Leam Howe (leamhowe1@gmail.com)

Abstract. Current operational snow mapping products struggle to accurately map patchy, late-season snow cover in challenging environments like Scotland, hindering hydrological and ecological model validation. We hypothesise that machine learning (ML) could better handle these variable conditions, including dirty and metamorphosed snow, cloud and atmospheric disturbance, varying illumination levels and shading.

5 To address this, we generated a novel Snow Cover Fraction (SCF) reference dataset for Sentinel-2 (10 m resolution) by aggregating manual and semi-automated "*pseudo-labels*" derived from 25 cm aerial surveys of Scotland. Using this dataset, we developed a Snow Patch U-Net (SPUN) by adapting a U-Net architecture with a regression head to predict SCF, utilising a combined Dice and L1 loss function. On our Scottish test set, SPUN successfully mapped challenging snow patches – even through cloud gaps and in topographic shadow – outperforming the operational Theia Let-It-Snow (LIS) product with a snow
10 segmentation F1 score of 0.737, a Mean Absolute Error (MAE) of 28.49%, and an RMSE of 42.16%.

Furthermore, by integrating pre-trained weights from satellite foundation models, our model demonstrated strong generalisation when applied to an independent global dataset, achieving a macro averages F1 Score of 0.9313 across the three classes (snow, cloud, background) and a snow-only F1 score of 0.8923.

This study demonstrates that ML has the potential to tackle persistent challenges in optical snow retrieval. We suggest that a
15 broader, community-driven effort to create a global, high-resolution training dataset could significantly advance snow mapping capabilities across optical satellite sensors.

1 Introduction

Seasonal snow cover is a critical component of the Earth's climate system that influences the global energy budget through the albedo effect and acts as a primary freshwater source for more than a sixth of the global population (Barnett et al., 2005; Flanner et al., 2011). Beyond the climatic and hydrological significance, snow cover dynamics play a fundamental role in ecological systems by modulating microbial activity and vegetation phenology, while also influencing the timing and distribution of natural hazards such as droughts, floods, and avalanches (Sturm et al., 2001; Wipf and Rixen, 2010; Livneh and Badger, 2020).

Snow has a high spatial extent and heterogeneity, varying from covering over half the northern hemisphere land area at maximum extent and retreating to just the highest latitudes and elevations during mid-summer (Estilow et al., 2015). On top

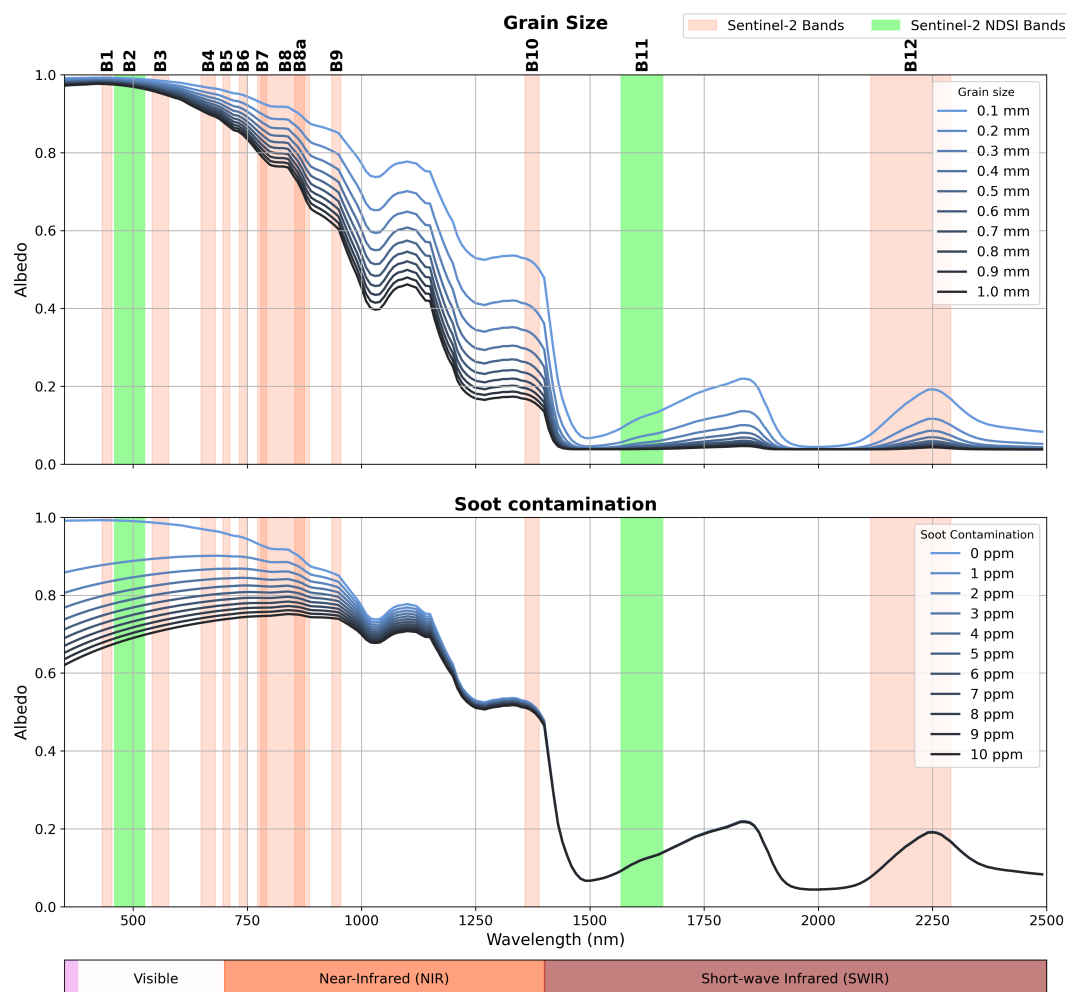


Figure 1. Influence of snow grain size and contamination on the reflective properties of snow. As snow gets older (blue → black lines), both grain size and contamination increase leading to a decrease in albedo in different parts of the spectrum. The Sentinel-2 bands are highlighted, with the specific bands used to calculate the Normalised Difference Snow Index (NDSI) (Hall et al., 1995) highlighted in green to emphasise how much the reflectivity of these bands is influenced by contamination and metamorphosis of the snow. Spectra simulated using the SnowTARTES model (Picard and Libois, 2024).

25 of this, snow often resides in remote, harsh and inaccessible environments, making remote sensing the only viable means to monitor snow consistently across relevant spatial and temporal scales. The Sentinel-2 (S2) mission, with its high spatial resolution (10 m), and frequent and regular revisit, has revolutionised our ability to monitor snow at a catchment scale (Drusch et al., 2012; Gascoin et al., 2019). This resolution is critical because snow cover variability in mountain terrain typically occurs at length scales of less than 100 m, meaning coarser heritage sensors like MODIS (500 m) often fail to capture the spatial

30 heterogeneity induced by complex topography (Gascoin et al., 2024).



However, despite these hardware advancements, operational monitoring remains hindered by algorithmic challenges. Traditional optical methods – primarily relying on spectral indices and thresholding – struggle to distinguish snow from cloud cover and suffer from significant errors in topographically complex terrain (Dozier, 1989; Hollstein et al., 2016; Gascoin et al., 2019). The performance of these algorithms is often overestimated because validation studies typically focus on clear-sky imagery, leaving the confusion between clouds and snow in operational contexts largely unaddressed (Gascoin et al., 2024).

Furthermore, the development and validation of snow algorithms have historically favoured regions with seasonally persistent snow cover, such as the Alps and the Rocky Mountains. Consequently, marginal and ephemeral snowpacks – typical of temperate maritime climates such as the Scottish Highlands – remain under-represented in the literature (López-Moreno et al., 2024). In these environments, snow is often patchy, wet, and subject to rapid melt-freeze cycles, substantially changing the spectral reflectance (see Figure 1). As the global climate warms, many regions are transitioning from stable seasonal snow regimes to these "marginal" regimes (Beniston et al., 2018; Bavay et al., 2009). In this shifting landscape, the ability to accurately map fractional snow cover in patchy conditions and within topographic shadows is no longer just a technical refinement, but a prerequisite for reliable future monitoring.

ML, and specifically Deep Learning (DL), offers a promising alternative to traditional spectral thresholding and unmixing methods. DL models, such as Convolutional Neural Networks (CNNs), can learn complex, non-linear relationships within spectral-spatial features, potentially resolving the confusion between cloud, snow, and shadow that limits current operational products. However, despite a growing body of literature exploring DL for S2 snow mapping (Wang et al., 2022; Lu et al., 2022; Demil et al., 2025), research has almost exclusively focused on binary Snow Cover Extent (SCE) classification. Predicting the sub-pixel Snow Cover Fraction (SCF) via DL remains largely unexplored, leaving a critical gap for hydrological applications that require continuous fractional estimates. Furthermore, a significant disparity remains between algorithmic development and user accessibility. Many studies do not publish their underlying datasets or code, creating a reproducibility barrier that discourages uptake by the wider hydrological and ecological communities. Consequently, concerted efforts to integrate these advanced data-driven methods into open-access, operational pipelines remain limited.

In this paper, we address these technical and open-science limitations by developing a novel, open-access training dataset specifically designed for the temperate, marginal snowpack of Scotland. By pairing 25 cm aerial photogrammetry with S2 imagery, we capture the spectral signature of patchy, dirty, and metamorphosed snow. This dataset allows us to investigate the so-called 'clear-sky bias' in current validation protocols, where valid snow observations are systematically excluded by overly conservative cloud masks.

Building on this dataset, we present SPUN (a Snow Patch U-Net), a deep learning framework adapted to regress fractional snow cover directly from S2 Level-1C top-of-atmosphere reflectance. To the best of our knowledge, this represents the first application of DL to go beyond binary classification and perform pixel-wise SCF regression from optical satellite imagery.

The specific objectives of this study are to:

1. **Develop a unique fractional snow cover dataset** for Scotland (a temperate maritime climate with a marginal snowpack).

This dataset is designed to stress-test algorithms outside their standard operating envelope, particularly addressing the



gap of validation in cloudy temperate regions, but also targeting small, dirty, and highly metamorphosed late-season snow that usually remains in topographically shaded regions.

2. **Train and evaluate a novel U-Net regression model optimised for fractional snow mapping (SPUN)**, specifically testing its ability to resolve patchy snow in an exceptionally cloudy and marginal snow environment.

3. **Benchmark the SPUN model** against the state-of-the-art operational Let-It-Snow product (Theia Snow Collection) (Gascoin et al., 2019) and assess its generalisability to global environments using a global labelled dataset of 40 images (Wang et al., 2022).

2 Methods

2.1 Study Region

The Scottish Highlands serve as a rigorous test bed for evaluating the robustness of optical snow mapping algorithms due to the temperate maritime climate and resulting ‘marginal’ snow regime. Unlike the seasonally stable snowpacks of the Alps or Arctic, the Scottish snowpack is ephemeral and highly dynamic, characterised by rapid accumulation and melt cycles, frequent cloud cover, and low solar angles. These factors combine to create a challenging environment for remote sensing, ideal for testing algorithmic performance under stress.

The region’s climate is dominated by prevailing south-westerly airflow derived from the North Atlantic. As these moist air masses encounter the complex topography of the Highlands, they create a distinct longitudinal climatic gradient from west to east:

1. **The Western Highlands and Islands:** Including the Isle of Skye, the Northwest Highlands, and the Nevis range, this zone is hyper-oceanic. Directly exposed to Atlantic weather systems, these areas experience intense orographic precipitation, high humidity, and rapid temperature fluctuations. This results in a highly transient snowpack that is often wet, slushy, and difficult to spectrally distinguish from the frequent low-level cloud and fog.

2. **The Eastern Highlands:** Moving inland towards the Cairngorms, the climate becomes comparatively drier due to the rain shadow effect of the western ranges. While this leeward position can induce localised warming during westerly airflows (the Foehn effect), the region generally experiences colder winter conditions due to its increased distance from the moderating Atlantic Ocean and its sustained high-elevation plateaus where snow cover is highly persistent. While still maritime, the snowpack here is heavily influenced by wind transport, with deep accumulations loading into lee slopes and corries, creating complex spatial heterogeneity driven by redistribution rather than just elevation.



Scotland study region and SCF dataset splits

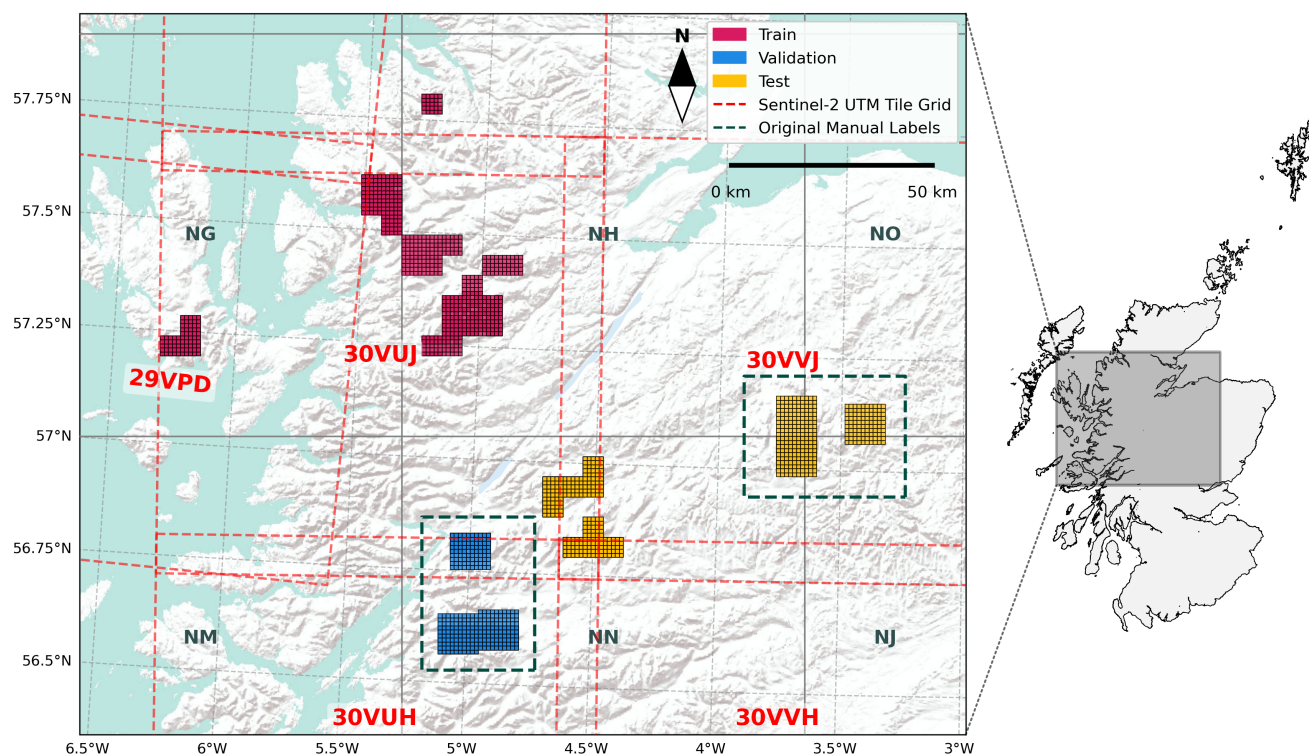


Figure 2. Map of study region and train/val/test data splits. These splits were done in an attempt to keep tiles from different S2 retrievals (red dashed lines) separate, whilst maintaining spatial separation and the balance of examples containing snow proportional within the train, validation and test sets. Map is oriented to the British National Grid (BNG) system with the 100 km grid boxes in grey and labelled with the appropriate identifier (e.g. NH); our labelled dataset is aligned to the 1 km grid boxes of the BNG and is seen as the small coloured boxes and the original manual labels are outlined with the dark dashed green lines, all other data was added with ‘pseudo-labelling’; the Sentinel-2 UTM tile grid is in dashed red lines; and the standard lat lon coordinates (WGS84) are labelled with faint dashed grey lines for reference. The ESRI World Terrain Base was using in creating this figure. Sources: Esri, USGS, NOAA | Powered by Esri.



2.2 Data – A Novel Scottish SCF Dataset

2.2.1 Satellite Data

We aimed to map snow at the highest-resolution possible using open-access satellite data, this being the S2 satellite sensor. This constellation is formed of two polar orbiting satellites which have a revisit time of 5 days and 13 spectral bands which range from 10 m to 60 m in ground sampling distance. In this study, we resample all bands to 10 m for consistent input resolution to the ML model and to enable the highest resolution output snow maps.

S2 data are made available at different processing levels. We use the S2 Level-1C Top-of-Atmosphere (TOA) data for two reasons: (1) so that data are comparable to other multi-spectral sensors; (2) because late-lying snow is found in topographically shaded corries where the atmospheric corrections involved in retrieving level-2 surface reflectance can be badly affected by shadow, and we wish to preserve the detail in these regions.

2.2.2 Data Annotation Pipeline

A labelled dataset is required for supervised ML. This often involves labour-intensive manual labelling of images through visual interpretation by experts. As explained in the introduction, distinguishing between snow and cloud in satellite imagery can be difficult even for an expert in the field, especially if images are heavily influenced by atmospheric affects. On top of this, part of this research was intended to investigate the application of DL for the regression task of predicting SCF, and it is impossible to manually attribute a SCF value to a satellite pixel reliably. Therefore, we opted to manually label very high-resolution (25 cm) aerial survey data, where the snow is easily identifiable, and then resample this to the resolution of S2 (10 m) to create a labelled dataset of SCF maps. Our pipeline is outlined in figure 3, and consists of three steps (1) label aerial data, (2) resample to 10 m SCF maps and match to a satellite image within 2 days of aerial retrieval, (3) mask cloud with established S2 cloud making algorithm.

GetMapping provides 25 cm aerial surveys of Scotland which are distributed as 1 km tiles matching the Ordnance Survey (OS) 1 km British National Grid (EPSG:27700). We labelled a subset of these data over the snowiest regions of Scotland, The Cairngorms, Ben Nevis, and Glen Coe (as seen by the dashed green line in Figure 2). This was still a fairly limited dataset of 849 image chips from 8 different S2 scenes. To increase the number of labels and avoid the time consuming labelling task, we trained a U-Net model to segment snow in the 25 cm aerial imagery; this U-Net model achieved a dice score of 0.9578 on an unseen test set, with a recall of 0.9510 and a high precision of 0.9859 on the binary snow classification task. We used this model to create new labels by predicting on unseen aerial imagery of other mountainous regions of Scotland. This process is known as "*pseudo-labelling*" and enabled us to increase our dataset from 849 to 2455 image chips from 37 different S2 scenes and 7 unique tiles across 7 different years, as shown in Figure 2.

Once we had these 1 km labelled aerial images, we matched each aerial image with the closest S2 scene within a 2-day window and resampled 25 cm binary snow labels to 10 m SCF labels. Since we are matching aerial data to satellite data there will be cloud in the satellite data that is not in the aerial data; we need to identify and label this as such. Cloud detection in S2 data is a well-studied field with numerous available algorithms, so we decided to use the state-of-the-art Google Cloud

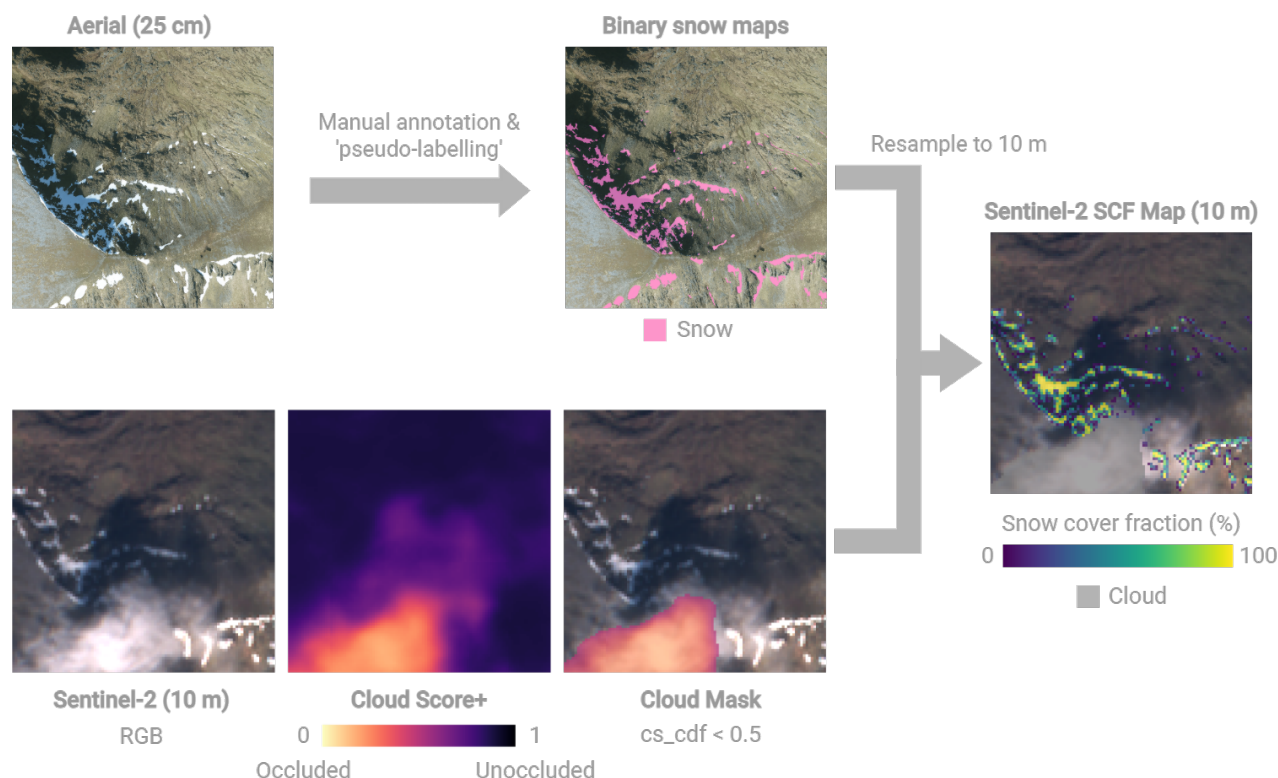


Figure 3. Labelling workflow to generate the SPUN Snow Cover Fraction (SCF) dataset. We attribute binary snow labels to 25 cm aerial data (© Getmapping Ltd) which is then resampled to generate 10 m SCF maps for Sentinel-2. Cloud is masked using the Google Cloud Score+ product thresholded at 0.5 to give a binary cloud mask.

Score+ (CS+) algorithm (Pasquarella et al., 2023). CS+ provides two scores, cs and cs_cdf , where 0 is more likely to be fully occluded and 1 is more likely to be clear. cs can be thought of as how similar the value is to a perfectly clear reference, whilst cs_cdf acts as a temporal percentile, indicating the relative rank of the current pixel's score compared to its entire historical distribution. cs_cdf has been shown to be less sensitive to low-magnitude spectral changes and terrain shadows; therefore, we used cs_cdf score in this study. While recommended CS+ thresholds range from 0.5 to 0.65, we applied a lenient threshold of 0.5 to classify clear-sky pixels. This approach deliberately preserves observations through cloud gaps and heavy atmospheric interference. Retaining these difficult viewing conditions allows us to rigorously test snow detection limits, addressing the 'clear-sky bias' in current validation protocols discussed by Gascoin et al. (2019).

The final dataset consists 2455 1 km $\times$ 1 km S2 SCF maps that can be used for supervised training of an ML algorithm. We split this dataset into train, validation, and test sets as shown in figure 2; these splits aim to preserve the balance between spatial separation, separation of S2 retrieval tiles, and the proportion of snow pixels within each split. The final splits contained 1052 training, 469 validation, and 934 test 1 km image chips. Further details on the dataset splits can be found in the appendix A.



2.3 SPUN model

2.3.1 Architecture

The U-Net architecture has become a staple for image segmentation tasks. Originally developed for medical image segmentation
140 (Ronneberger et al., 2015), this architecture has proven applicable to numerous tasks and, despite the rapid development of
the field and a shift towards transformers, it still proves to be amongst the best methods for image segmentation with a limited
training set (Marsocci et al., 2025). Here, we have a limited training set and a U-Net architect enables us to include both spectral
information from the 13 bands of S2 as well as spatial information from the use of convolutions to extract important spatial
features from the satellite imagery. Simply put, this method considers spatial as well as spectral features when identifying snow
145 which goes beyond pixel-based classification methods which rely solely on the spectral information.

However, estimating SCF requires more than just binary segmentation; it is a regression problem. To address this, we adapted
a ResNet50 architecture as our model’s encoder (backbone). This deep residual network excels at feature extraction and helps
mitigate the vanishing gradient problem in deep architectures. The final decoder output is then fed into a regression head to
predict the continuous percentage of snow within each pixel (0–100%). Figure 4 outlines the architecture.

150 2.3.2 Optimisation and Training Settings

To simultaneously optimise for both categorical detection (where is the snow) and numerical precision (what percentage of
snow is in a pixel), we employ a hybrid loss function:

$$L_{total} = \lambda L_{dice} + (1 - \lambda) L_{reg} \quad (1)$$

This combined approach uses a Soft Dice loss (L_{dice}) to ensure accurate spatial overlap of snow patches, and a regression
155 loss (L_{reg}) to minimise the error in SCF percentage. By tuning the weighting parameter, λ , on the validation dataset, we
achieve a model that maximises detection accuracy (high Dice score) while maintaining low SCF estimation error.

The convolutional layers and 2×2 max-pooling operations within the U-Net architecture require input dimensions to be
powers of two to maintain integer feature map sizes. It is common practice to pad images to match these dimensions, therefore
we pad our original 100×100 pixel ($1 \text{ km} \times 1 \text{ km}$ at 10 m resolution) image chips to 128×128 pixels before inputting them
160 into the network.

SPUN was developed using the Segmentation Models PyTorch (SMP) and TorchGeo extensions for PyTorch (Stewart et al.,
2022; Paszke et al., 2019). We explored a variety of backbones and pretrained weights and settled on a Resnet50 architecture
with S2 Dino weights from the SSL4EO-S12 framework (He et al., 2015; Wang et al., 2023). This self-supervised pretraining
on satellite imagery provides a superior starting point for remote sensing tasks compared to standard ImageNet or randomised
165 weights (Wang et al., 2023). Finally, to address the severe class imbalance in our dataset – where snow-covered pixels are
vastly outnumbered by cloud or background pixels (see appendix A) – we used a per-pixel weighted loss function to ensure the
model prioritises the accurate classification of the rare snow class.

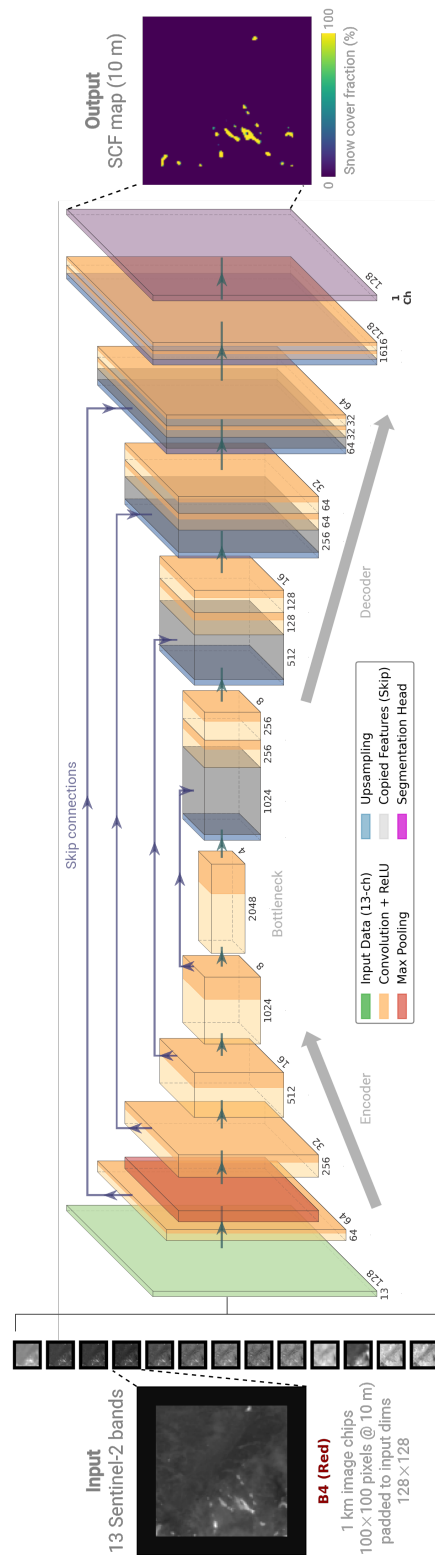


Figure 4. Diagram of the Snow Patch U-Net (SPUN) architecture. The model ingests a 13-band Sentinel-2 satellite image patch (100×100 pixels), which is padded to create a $128 \times 128 \times 13$ input tensor, and outputs a $128 \times 128 \times 1$ Snow Cover Fraction (SCF) mask. U-Net architectures utilise an 'encoder' (left half of diagram) that applies convolutional filters to extract spatial patterns, periodically downsampling the data to capture broader contextual features. The 'decoder' (right half of diagram) then upsamples these features to reconstruct the spatial resolution. Skip connections (purple arrows) transfer high-resolution spatial details directly from the encoder to the decoder, allowing the model to precisely localise snow boundaries. The network terminates with a Sigmoid activation and is trained using a combined Dice and L1 loss function. This figure was created using the PlotNeuralNet library (<https://github.com/HarisIqbal88/PlotNeuralNet>).



During training, the network weights were updated using the Adam optimiser (Kingma and Ba, 2017). To determine the most effective configuration for this training process, hyperparameter tuning was conducted using the Optuna framework (Akiba et al., 2019). We optimised the model to simultaneously minimise RMSE on snow and maximise the Dice score, identifying the Pareto optimal trade-offs between prediction precision and segmentation accuracy. The search space spanned 100 trials per study and evaluated various pretrained backbones, including ImageNet, DINO, MoCo, DeCUR, and SoftCon, alongside a randomly initialised baseline. Within each study, we tuned the learning rate (1×10^{-5} to 1×10^{-2}), batch size (8, 16, 32), and regression loss functions (L1, MSE, Huber). Task-specific parameters were also optimised, specifically: the dice loss weight, which determines the relative importance of spatial overlap (0.1–0.9); the snow weight, which defines the focus placed on the minority snow class (1.5–20.0); and the decision boundary, the fractional threshold used to binarise the maps for spatial evaluation (0.01–0.5). The optimum parameters found were DINO, learning rate of 0.000118, batch size of 16, a Huber regression loss function, dice loss weight of 0.166421, no snow weighting, a snow decision boundary of 0.399728.

2.4 Experimental Design

To rigorously assess the performance of SPUN, we conducted a comparative analysis against operational baselines and existing deep learning approaches.

Table 1. Summary of Evaluation Tiers. Overview of the testing scenarios used to validate model performance across different spatial resolutions and atmospheric conditions.

Tier	Scenario	Description
Tier 1	Intrinsic (10 m)	Direct evaluation against labels at the native resolution of the SPUN model.
Tier 2a	Clear-Sky Control (20 m)	Comparison against the operational LIS product under strictly cloud-free conditions.
Tier 2b	All-Sky Reality (20 m)	Comparison against the operational LIS product under typical Scotland weather conditions, assessing the impact of clouds.

2.4.1 Baseline Algorithms

We benchmarked SPUN against three categories of methods:

- **NDSI > 0.4 (binary output @ 10 m)** a simple threshold on NDSI to provide a baseline.
- **Operational Baseline (Theia/LIS) (fractional output @ 20 m):** The *Let-it-Snow* algorithm (Gascoin et al., 2019), representing state-of-the-art spectral thresholding.



- **Other Deep Learning Methods (3-class categorical output @ 10 m):** The CNN architecture proposed by Wang et al. (2022) was made openly available so we applied this to our Scottish dataset to compare performance. For the global comparison we also compared to Ma et al. (2023) and Demil et al. (2025) which are developments on Wang et al. (2022) but do not make their code available so we could not compare their performance on our Scotland dataset.

Table 1 summarises the 3 different scenarios that we compare the snow mapping performance across. 10 m products are compared at their native resolution (Tier 1: Intrinsic), and then resampled to 20 m resolution for comparison with the LIS product (Tier 2). Although NDSI and Wang et al. (2022) are binary/categorical classifiers, we evaluate their effective fractional performance when aggregated to the operational 20 m grid, this allows us to quantify the fractional error introduced by binary quantisation at the native sensor resolution. At the Tier 2 (20 m) analysis level, we further break down the comparison into ‘Clear-Sky’ (Tier 2a) and ‘All-Sky’ (Tier 2B) to explore the influence of cloud masking.

2.4.2 Performance Metrics

The task of predicting SCF from satellite data presents a dual challenge: correctly identifying the spatial extent of the snowpack (a segmentation problem) and accurately quantifying the snow fraction within each pixel (a regression problem). To comprehensively evaluate the model’s performance on both fronts, we employ distinct metrics for each task and also add an aggregate chip-scale accuracy to assess hydrological utility.

- **Regression Accuracy (RMSE & MAE):** Pixel-level precision was evaluated using Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). Crucially, to account for the zero-inflated nature of the dataset – where the vast majority of pixels are snow-free – we computed both "All-Class" and "Snow-Only" metrics. Due to the sparsity of snow in this dataset, we consider the "Snow-Only" metrics to be the definitive indicator of regression performance. The "Snow-Only" metrics are calculated exclusively on the *union* of all true and predicted snow pixels ($y_{true} \geq 1\%$ or $y_{pred} \geq 1\%$), ensuring both omission (missed snow) and commission errors (false alarms) are captured. This approach prevents the massive number of correctly identified background pixels (where both prediction and label are zero) from artificially lowering the overall error statistics, while still strictly penalising the model for hallucinating snow on bare ground.
- **Segmentation Accuracy (Dice Score)** To assess spatial accuracy, continuous predictions were binarised using a 50% SCF threshold. We calculated the Dice Similarity Coefficient (F1 Score) to measure the spatial overlap between the predicted snow mask and the label. This enabled further validation and comparison against datasets and models that only deliver a SCE product.
- **Snow Cover Area (SCA)** To assess the model’s utility for regional snow estimation – specifically its ability to track the true magnitude of the snowpack extent – we compared the Total Snow Covered Area (SCA) derived from each product across the 1 km test chips. Unlike pixel-wise metrics, this aggregate analysis reveals whether local spatial errors compound or cancel out over a defined area. We calculated SCA by summing the total snow area (in km^2) at the native resolution of each product, filtering exclusively for "snow-relevant" chips (i.e., where either the reference label or the



model prediction was > 0). This approach allows us to accurately quantify total area bias at the chip scale without
skewing the results using entirely snow-free baseline imagery.

2.4.3 From Scotland to Global Generalisability

We assessed model performance using a two-stage validation protocol. First, to evaluate fractional snow retrieval in a marginal environment, we utilised the held-out test split of our Scottish dataset. To ensure rigorous independence, this test set was spatially separate from the training and validation partitions (see Fig. 2), ensuring the model was evaluated on unseen terrain. The Scotland performance metrics for SPUN and the baseline methods were derived exclusively from this independent subset.

Second, to assess the generalisability of our model beyond the training domain, we applied SPUN to the independent global dataset curated by Wang et al. (2022). This dataset consists of 40 manually labelled S2 scenes distributed across six continents and diverse biomes as seen in figure C1. As this external dataset provides binary labels, it serves as a test of the model's ability to detect snow extent in novel environments without site-specific retraining. We compare the operational SPUN pipeline (incorporating CloudScore+ masking to provide the cloud class) against three deep learning models trained specifically on this global dataset: the U-Net baseline (Wang et al., 2022), UCTNet (Ma et al., 2023), and mDeeplabv3+ (Demil et al., 2025). Unlike these baselines, which are trained end-to-end to classify clouds, SPUN is trained to classify snow, ignoring clouds, and a cloud mask can be applied independently.

3 Results

3.1 Performance on our Scotland dataset

SPUN is able to distinguish snow from cloud and other land cover even under heavy atmospheric influence and through gaps in clouds and in shadow. Figure 5 visualises seven example chips from the test dataset of SPUN compared to other methods, including NDSI as a baseline. We see LIS underpredicting snow, and most of the time masking it out completely as cloud. NDSI shows the known issues of struggling to detect dirty and metamorphosed snow and misclassifying water, but shows its clear benefits when handling clouds. Wang et al. (2022) shows limited ability to detect snow in the conditions presented in this dataset, likely because the viewing conditions and snow state in Scotland are substantially different to what this U-Net algorithm was trained on.

3.1.1 SCF Regression Accuracy

Table 2 presents the error metrics for SPUN compared against the operational LIS product and baselines. We report both "All-Class" and "Snow-Only" metrics to distinguish global dataset statistics from performance on the target class. All-Class metrics (MAE and RMSE) remain consistently low across all methods ($\leq 0.35\%$), however, these values are heavily influenced by the dominant snow-free class. Consequently, we consider the "Snow-Only" metrics to be the primary indicator of regression performance.

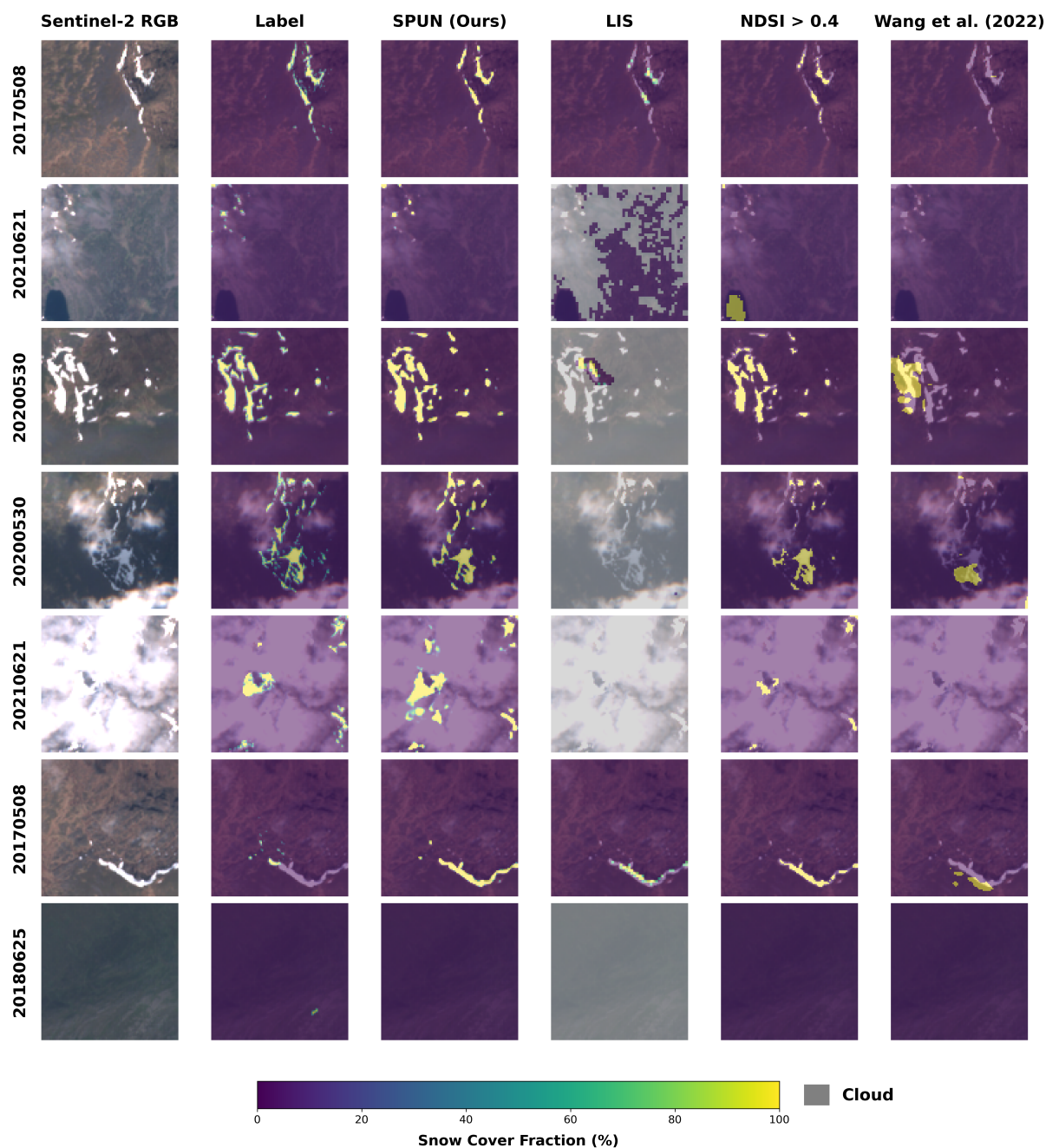


Figure 5. Visual examples of 1 km Sentinel-2 image chips from test dataset. Each row shows a different image chip, each column shows a different snow mapping method. There are 6 columns, from left to right these are the Sentinel-2 RGB composite, the labelled data, SPUN (our) output, Let-It-Snow (LIS) operational algorithm, NDSI > 0.4, and the Wang et al. (2022) output. Row 1 is an example of well illuminated snow patches where algorithms are expected to do well; row 2 are small patches of snow with minor cloud influence and a lake; row 3 is an image with mild atmospheric distortion which is classified as cloud by LIS; row 4 show snow in cloud shadow; row 5 heavy cloud cover with views of snow through gaps; row 6 and 7 show errors in the labels which are overcome by SPUN.



Intrinsic Performance (Tier 1): At the native 10 m resolution, the error metrics highlight the distinction between binary
250 and continuous outputs. The binary baselines, NDSI and Wang et al. (2022), exhibit high Snow-Only MAE values (61.70%
and 54.70%, respectively), reflecting the quantisation error inherent in representing fractional snow cover with discrete classes.
In contrast, SPUN achieves a substantially lower MAE of 28.49% and RMSE of 42.16%, indicating that its regression output
better approximates the continuous distribution of snow cover values at this resolution.

Operational Performance (Tier 2): When resampled to the 20 m operational grid, SPUN maintains a distinct performance
255 advantage. Under Clear-Sky conditions (Tier 2a), SPUN reduces the Mean Absolute Error (MAE) by 4.2 percentage points
relative to LIS (19.10% vs. 23.26%) – an 18% relative reduction in error. However, SPUN's true advantage emerges under All-
Sky conditions (Tier 2b) when optical visibility is compromised. LIS suffers a sharp increase in error (MAE 37.78%) driven by a
severe negative 'Snow-Only Bias' (-37.23%), as cloudy areas are strictly masked and underlying snow is systematically missed.
By contrast, SPUN effectively infers fractional coverage, maintaining robust performance (MAE 18.81%). Furthermore, SPUN
260 achieves this with only a slight positive bias (+6.10%), proving it successfully recovers hidden snow without triggering
excessive false positives.

The results for NDSI and Wang et al. (2022) must be interpreted with their structural differences in mind. Unlike SPUN
and LIS, these are binary/categorical classifiers at their native 10 m resolution. When aggregated to the 20 m comparison grid,
their effective "fractional" precision is quantised to steps of 25% (0, 25, 50, 75, 100), inherently limiting their ability to resolve
265 fine-scale gradients in snow cover. Consequently, their higher error rates (e.g., Wang et al. Clear-Sky MAE 28.91%) should be
viewed not solely as model inaccuracy, but as a demonstration of the limitations of binary segmentation. Wang et al. shares the
conservative negative bias of LIS (-22.41%), while NDSI suffers from overestimation (Bias +16.88%) due to water confusion.

3.1.2 Binary Snow Cover Extent (SCE)

Table 3 details the binary classification performance, where fractional predictions (SPUN and LIS) are thresholded at 50% to
270 determine snow presence.

Intrinsic Performance (Tier 1): At the native 10 m resolution, SPUN demonstrates robust detection capability under All-
Sky conditions, achieving an F1 Score of 0.737. This significantly outperforms standard baselines; NDSI (>0.4) achieves an
F1 Score of only 0.313, hampered by low precision (0.246), while the ML baseline of Wang et al. (2022) fails to generalise to
the domain (F1 0.191).

Operational Performance (Tier 2): Under Clear-Sky conditions at 20 m (Tier 2a), the trade-offs between the methods
275 become apparent. The operational LIS product achieves the highest precision (0.562) of the group. While this indicates a
relative preference for minimising false positives compared to the other methods, the moderate absolute value highlights the
difficulty of spectral classification of late-lying snow in the Scottish Highlands. Furthermore, this marginal gain in precision
comes at the cost of sensitivity; LIS recall is lower (0.439), meaning it omits more than half of the snow present even when
280 visibility is perfect. In contrast, SPUN prioritises retrieval completeness, achieving a balanced F1 Score (0.606) and a Recall
of 0.789 – nearly double that of the operational baseline – albeit with a reduction in Precision (0.492).



Table 2. Snow Cover Fraction (SCF) Error Analysis. Comparison of error metrics across resolutions and scenarios. All-Class Metrics include non-snow pixels, which deflate error values due to class imbalance (zero-inflation). Snow-Only Metrics are calculated exclusively on the union of snow pixels ($\geq 1\%$), capturing the true error magnitude during snow events. Bold highlights the best performing method based on metric. Arrows indicate the optimal direction for each metric ($\downarrow$ indicates lower is better, $\rightarrow 0$ indicates closer to zero is better).

Scenario	Model	All-Class Metrics			Snow-Only Metrics		
		MAE ($\downarrow$)	RMSE ($\downarrow$)	Bias ($\rightarrow 0$)	MAE ($\downarrow$)	RMSE ($\downarrow$)	Bias ($\rightarrow 0$)
Tier 1: Intrinsic (10 m, All-Sky)	SPUN	0.12	2.69	+0.08	28.49	42.16	+8.36
	NDSI (> 0.4)	0.35	5.61	+0.22	61.70	74.24	+22.31
	Wang et al.	0.21	4.15	-0.31	54.70	67.00	-31.33
Tier 2a: Control (20 m, Clear-Sky)	SPUN	0.04	1.37	+0.01	19.10	28.98	+2.74
	LIS (Ops)	0.05	1.51	-0.03	23.26	33.30	-14.06
	NDSI (> 0.4)	0.47	6.39	+0.40	65.67	75.64	+55.84
	Wang et al.	0.06	1.90	-0.04	28.91	41.17	-16.38
Tier 2b: Reality (20 m, All-Sky)	SPUN	0.11	1.99	+0.04	18.81	26.58	+6.10
	LIS (Ops)	0.19	3.54	-0.19	37.78	50.09	-37.23
	NDSI (> 0.4)	0.35	5.16	+0.13	45.69	59.32	+16.88
	Wang et al.	0.21	3.74	-0.12	38.70	51.07	-22.41

Table 3. Binary Snow Detection Performance ($> 50\%$ Threshold). Comparison of classification capabilities at the native SPUN 10 m resolution (Tier 1) and native LIS 20 m resolution (Tier 2), under Clear-Sky (excluding cloud pixels using LIS strict cloud masking) and under All-Sky conditions. Bold highlights the best performing method based on metric. Arrows indicate the optimal direction for each metric ($\uparrow$ indicates higher is better).

Scenario	Model	Recall ($\uparrow$)	Precision ($\uparrow$)	F1 Score ($\uparrow$)
Tier 1: Intrinsic (10 m, All-Sky)	SPUN (10 m)	0.821	0.668	0.737
	NDSI (> 0.4)	0.432	0.246	0.313
	Wang et al.	0.132	0.347	0.191
Tier 2a: Control (20 m, Clear-Sky)	SPUN (20 m)	0.789	0.492	0.606
	LIS (Ops)	0.439	0.562	0.493
	NDSI (> 0.4)	0.333	0.028	0.051
	Wang et al.	0.035	0.093	0.051
Tier 2b: Reality (20 m, All-Sky)	SPUN (20 m)	0.874	0.652	0.747
	LIS (Ops)	0.013	0.562	0.026
	NDSI (> 0.4)	0.396	0.232	0.293
	Wang et al.	0.134	0.354	0.194



Table 4. Snow Cover Area (SCA) Retrieval. Evaluation of aggregate snow area (km^2) calculated at the native resolution of each product across snow-relevant image chips (where label or Model > 0). Clear-Sky Control ($N = 33$) isolates model sensitivity in ideal conditions. All-Sky Reality ($N = 165$) assesses operational reliability in typical weather. "Area Deviation" denotes the percentage difference in total area relative to the label reference. Bold highlights the best performing method based on metric.

Scenario	Model	R^2 Score	Total Area (km^2)	Area Deviation
Clear-Sky Control ($N=33$)	<i>Label</i>	<i>1.000</i>	<i>0.068</i>	<i>Reference</i>
	SPUN (10 m)	0.495	0.077	+12.3%
	LIS (20 m)	0.270	0.029	-57.7%
	NDSI (> 0.4)	-76.887	0.613	+795.6%
	Wang et al.	-0.149	0.021	-70.0%
All-Sky Reality ($N=165$)	<i>Label</i>	<i>1.000</i>	<i>1.763</i>	<i>Reference</i>
	SPUN (10 m)	0.934	2.083	+18.2%
	LIS (20 m)	-0.238	0.029	-98.4%
	NDSI (> 0.4)	-4.811	2.954	+67.6%
	Wang et al.	0.475	0.641	-63.6%

Opening the analysis to operational "All-Sky" conditions (Tier 2b), the divergence in model performance illustrates the limitation of current optical products. LIS performance effectively collapses (Recall 0.013, F1 0.026) because its strict cloud mask flags the vast majority of the test data as invalid. While this preserves data integrity by avoiding cloud contamination, it results in near-total data loss during typical Scottish viewing conditions. In contrast, SPUN maintains high robustness in this operational reality, achieving a Recall of 0.874 and Precision of 0.652. This suggests that SPUN can effectively infer snow presence even in pixels that strict masking algorithms discard, transforming "missing data" into usable information.

The binary baselines exhibit distinct failure modes. NDSI (> 0.4) suffers from extreme commission errors in the Tier 2a Clear-Sky scenario, with a Precision of just 0.028. Wang et al. (2022) fails to generalise to this dataset of late-lying snow in the Scottish Highlands, likely due to the snow possessing significantly different spectral properties compared to the model's training distribution. Visual inspection of the output (Figure 5) reveals that this model sometimes identifies snow in the vicinity of patches but fails to delineate the boundaries correctly. This suggests that the model has not learned the robust spectral features of snow, but rather to segment roughly appropriate regions based on spatial context, which proves insufficient for the complex, fragmented snow cover of this domain.

3.1.3 Snow Cover Area (SCA)

Table 4 reports the aggregate retrieval of SCA for each model.

Clear-Sky Control: To isolate algorithmic sensitivity to fractional snow, we first evaluated a Clear-Sky control scenario. Because strict cloud filtering reduces the sample size to 33 chips, these results serve as a mechanistic demonstration of



detection capabilities rather than an exhaustive climatological validation. Despite this limited sample, a clear signal emerges:
300 the operational LIS product severely underestimates total snow area by 57.7% (0.029 km^2 vs. the 0.068 km^2 reference). This indicates that the rigid spectral thresholds employed by LIS routinely omit the thin, patchy snow characteristic of this dataset. In contrast, SPUN recovers a total area of 0.077 km^2 (a slight overestimation of +12.3%) and achieving a notably higher spatial correlation ($R^2 = 0.495$) compared to LIS ($R^2 = 0.270$).

All-Sky Reality: The performance gap widens significantly in the All-Sky scenario, where the larger sample size (N=165)
305 provides a more reliable basis for evaluation. SPUN demonstrates strong predictive skill, achieving a near-linear correlation with label area ($R^2=0.934$). Its total area deviation remains stable at +18.2%, indicating that the model captures the majority of the snow extent with a consistent positive bias. Conversely, on this dataset, LIS is severely limited by cloud masking. The model achieves a negative R^2 score (-0.238), indicating no predictive skill relative to the mean chip SCA. Quantitatively, LIS recovers only 1.6% of the total reference area (0.029 km^2 vs 1.763 km^2), representing a -98.4% loss of signal.

310 **Baseline Areal Errors:** The binary baselines exhibit extreme areal deviations. NDSI (>0.4) overestimates area by +795.6% in clear conditions and +67.6% in all-sky conditions, confirming that simple spectral thresholding is prone to large-scale false positives in this environment. Wang et al. (2022) mirrors the conservative behaviour of LIS, systematically underestimating area by -63.6% to -70.0% across scenarios.

3.2 Performance on Global Dataset

315 To assess the generalisability of SPUN beyond the Scottish training domain, we evaluated the model on the global semantic segmentation dataset provided by Wang et al. (2022). This dataset includes diverse snow and cloud conditions that differ spectrally and spatially from that found in the Scotland. Figure C1 shows the locations of the 40 images in this dataset.

3.2.1 Global Benchmark Comparison

Table 5 presents the macro-averaged metrics. SPUN proves highly competitive, achieving a macro-averaged F1 Score of 0.931
320 and mIoU of 0.874. This performance is effectively at parity with the original Wang et al. baseline (F1 0.939) and remains comparable to recent transformer-based architectures like UCTNet (F1 0.950). We were only able to retrieve LIS snowmaps for 4 of the 6 test images because the LIS product relies on MAJA processed Level-2 data (Hagolle et al., 2017), which does not work if the individual scene or previous scenes are over a certain cloud threshold. However, we ran the LIS snow mapping algorithm on the 4 images we were able to retrieve and the results are listed in Table 5. SPUN outperformed LIS across all
325 metrics used in this assessment. We also investigated if this result was due to differences in clouds masking, so again computed "snow-only" scores and these are reported in the appendix Table C2, but the results were consistent.



Table 5. Global Benchmark (Macro-Averages). Comparison of the operational SPUN pipeline (SPUN snow predictions combined with CloudScore+ cloud mask) to the operational Let-It-Snow (LIS) product and segmentation models from 3 previous studies that were trained on the Wang et al. (2022) dataset (Wang et al., 2022; Ma et al., 2023; Demil et al., 2025). Metrics represent the average of Snow, Cloud, and Background class scores. Metrics for Ma et al. and Demil et al. are quoted directly from their respective publications, as their original models were unavailable for reproduction.

Metric	LIS* (Ops)	Wang et al. (2022)	Ma et al. [†] (2023)	Demil et al. [†] (2025)	SPUN (Ours) (Pipeline)
Precision	0.830	0.945	0.959	0.999	0.941
Recall	0.814	0.934	0.943	0.977	0.924
F1 Score	0.803	0.939	0.950	0.980	0.931
mIoU	0.672	0.886	0.906	0.911	0.874

* Evaluated on a 4-tile subset where LIS was retrievable.

[†] Metrics quoted directly from the original manuscripts.

Table 6. Ablation Study (Snow-Only Metrics). Comparison of the Raw Model against the Operational Pipeline (Masked). "Full Dataset" ($N = 40$) and "Test Subset" ($N = 6$). False Positive Rate (FPR) acts as a proxy for cloud confusion.

Configuration	Dataset	Precision	Recall	F1 Score	FPR
Raw SPUN	Subset ($N = 6$)	0.723	0.837	0.776	9.5%
(No Mask)	Full ($N = 40$)	0.651	0.919	0.762	24.3%
Operational	Subset ($N = 6$)	0.956	0.837	0.892	1.2%
(With Mask)	Full ($N = 40$)	0.883	0.918	0.900	5.8%

3.2.2 Impact of Cloud Masking (Ablation)

To disentangle the contribution of the cloud mask from the intrinsic capability of the SPUN model, we conducted an ablation study analysing "Snow-Only" performance with and without the masking module (Table 6). In this analysis, cloud pixels in the label are treated as "Non-Snow", penalising any cloud-snow confusion.

The results highlight the necessity of the masking module for global application. The raw (unmasked) SPUN model exhibits a False Positive Rate (FPR) of 24.3% on the full dataset ($N=40$), indicating spectral confusion between the clouds in this dataset and the representation of clouds that SPUN has learnt from Scotland. However, the operational mask effectively suppresses these errors, reducing the FPR to 5.8% while preserving high recall (0.918). This confirms that the precision reported in the benchmark comparison is achieved through the modular combination of SPUN's high sensitivity and CloudScore+'s effective cloud masking.

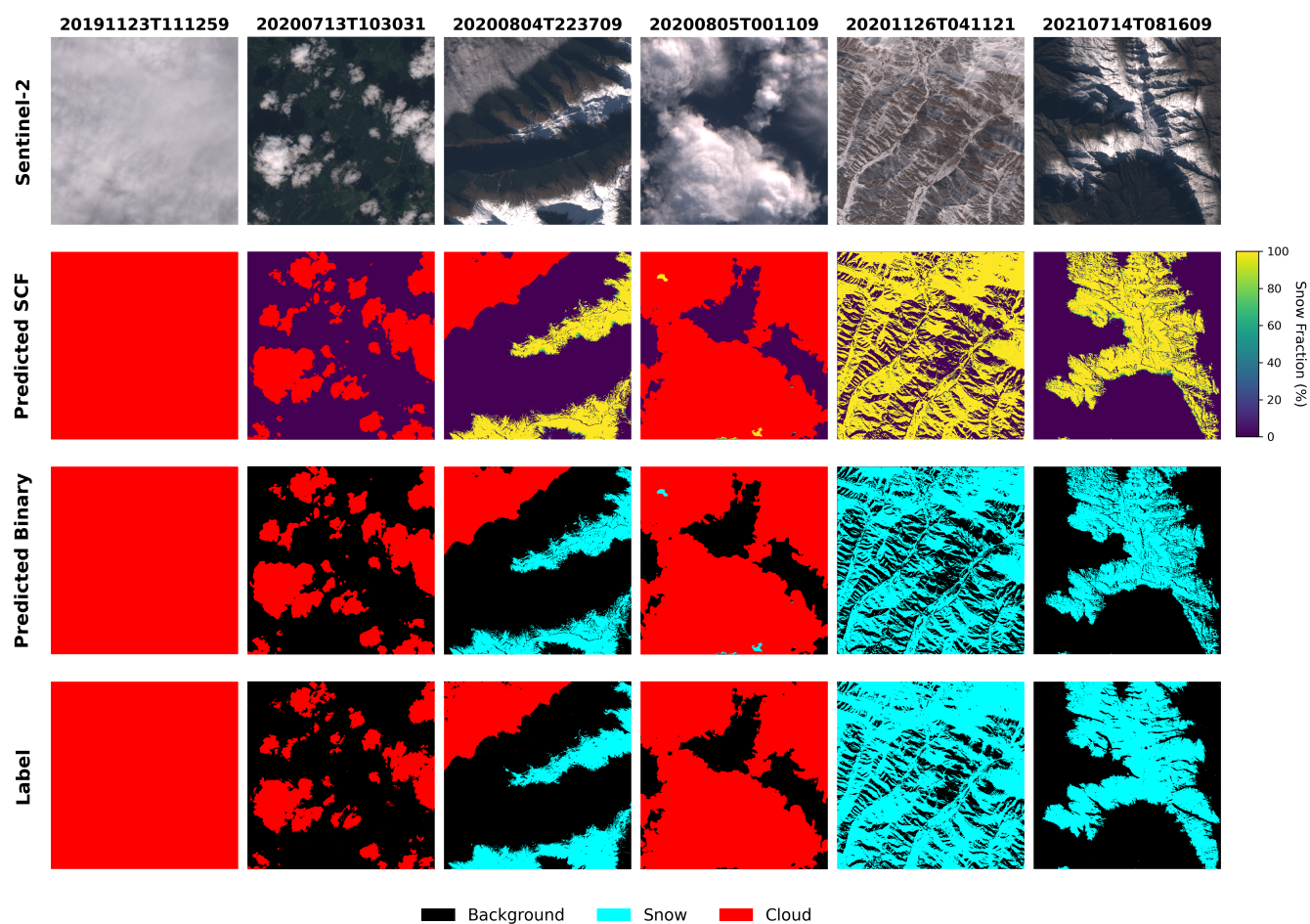


Figure 6. Visual Assessment of Global Predictions. Comparison of SPUN model outputs against labels for the six images in the test subset from Wang et al. (2022). Top to bottom: (1) Sentinel-2 RGB Imagery (gamma-enhanced); (2) Predicted Snow Cover Fraction (Viridis colormap, clouds masked in red); (3) Binary Prediction (>50% SCF classified as Snow, Cyan); (4) Label Masks (Cyan=Snow, Black=Background, Red=Cloud).

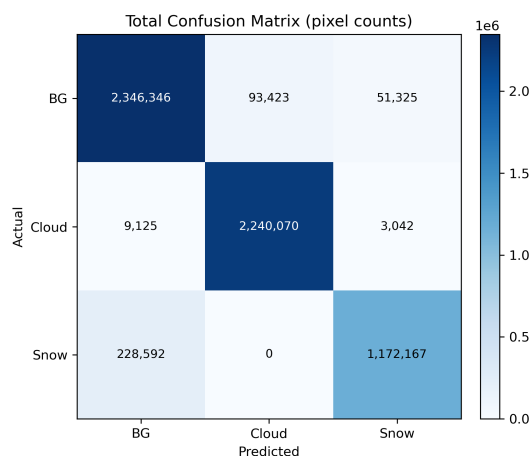


Figure 7. Confusion Matrix (Global Test Subset). Pixel-wise classification counts for the final masked SPUN predictions against labels (N=6). The matrix highlights that after cloud masking, the primary error mode is the omission of snow (Snow classified as Background).

3.2.3 Qualitative Assessment and Error Analysis

Having established the role of the mask, we examine the residual errors in the final operational product. Figure 6 presents the visual results for the test subset, displaying the RGB imagery, the continuous SCF prediction, the thresholded binary snow map (SCF>50%), and the label mask. Qualitatively, SPUN correctly identifies the majority of snow cover with predicted binary snow maps looking very similar to the labels.

The distribution of residual errors is quantified in the confusion matrix (Figure 7). In the final masked output, cloud confusion is negligible, with only 0.13% of cloud pixels misclassified as snow (3,042 pixels). Instead, the primary source of error is omission: the model misclassifies approximately 16.3% of true snow pixels as background. This conservative behaviour results in a lower recall (83.7%) compared to precision (95.6%), suggesting that while the operational pipeline rarely hallucinates snow (thanks to the mask), it does miss some snow viewing conditions which are likely under-represented in the Scottish training data. The confusion matrix when SPUN is applied to all 40 images in the dataset is provided in figure C4 and is more representative of SPUN's actual performance, showing the tendency to overestimate snow which is consistent with that observed on our Scotland dataset.

4 Discussion

The results of this study demonstrate that deep learning architectures, specifically the SPUN model, can offer a significant benefit over traditional spectral thresholding methods for fractional snow cover mapping. Several previous studies have used U-Net architectures to perform binary classifications of SCE (Wang et al., 2022; Ma et al., 2023; Demil et al., 2025), but to our knowledge, SPUN is the first study to extend this approach to regress continuous percentage SCF, an important input for



355 downstream hydrological modelling and water resource quantification in complex terrain. Rather than attempting to create
a universal global model, this study focused on correctly identifying small, patchy, and marginal snow in the landscape
– a notoriously difficult task for operational optical satellites. However, we took this one step further and have shown the
generalisability of this model as a proof of concept that these techniques are valuable for global snow cover mapping.

4.1 Performance in Marginal Environments (The Scottish Dataset)

360 The evaluation across our dataset of patchy marginal snow in Scotland highlights the limitations of current snow mapping
algorithms in maritime environments characterised by substantial cloud cover and fragmented/marginal snowpacks. We emphasise
that this dataset is a harsh test of snow mapping algorithms with late-lying snow spectra being quite different to that of fresh
snow, as seen in Figure 1.

4.1.1 Leveraging Spatial Context to Overcome Spectral Constraints

365 Simple spectral indices, such as the NDSI (>0.4) baseline, proved highly inadequate, overestimating snow area by nearly 800%
in clear-sky conditions (Table 4). This massive overestimation is symptomatic of environments where bare rock, water bodies,
and varying illumination mimic the spectral signature of snow. Conversely, the operational LIS product – which employs strict,
multi-spectral conservative thresholds to avoid these exact false positives – suffers from severe omission errors, underestimating
the true snow area by 57.7% even in ideal, cloud-free conditions. SPUN navigates this compromise effectively by leveraging
370 spatial context: by utilising a convolutional architecture, SPUN learns the textural features of the snowpack (e.g., the shape
of late-lying snow patches against specific topographic features) in addition to the spectral information. This allows SPUN to
identify fractional and dirty snow that falls outside the rigid spectral thresholds of LIS.

SPUN significantly outperforms all tested algorithms on the Scotland dataset. While the LIS algorithm exhibits a *precise*
method for predicting SCF values on the few pixels it confidently identifies, major omission errors mean it is consistently
375 outperformed by SPUN across both the SCF error metrics (Table 2) and binary SCE metrics (Table 3) for Scotland. Visual
inspection reveals that SPUN overestimates SCF at the boundaries of snow patches, generating sharper, near-100% predictions
in edge pixels rather than a smooth fractional gradient (as seen in Figure 5). This behaviour is a direct consequence of the
training data distribution and the hybrid loss function. Fractional snow cover inherently exhibits a bimodal class imbalance,
where pure background (0%) and pure snow (100%) vastly outnumber intermediate mixed pixels located exclusively at patch
380 perimeters. Consequently, the model learns a strong prior for extreme values. This polarisation is further reinforced by the
inclusion of the Dice loss, which adds incentive to put snow in the correct spatial location (SCE) over achieving the perfect
sub-pixel fraction (SCF). However, for hydrological utility, this slight boundary overestimation is preferable to the severe
omission errors seen in the operational products.



4.1.2 Operational Resilience in All-Sky Conditions

385 A recent review of methods for remote sensing of mountain snow (Gascoin et al., 2024) highlighted that studies often overestimate snow mapping abilities because validation scenarios focus exclusively on clear-sky imagery, thereby masking operational omission errors. We addressed this issue by comparing a Clear-Sky (Tier 2a) scenario to an All-Sky (Tier 2b) scenario (Table 2 and 3). The All-Sky (Tier 2b) scenario is where SPUN shows its strength. In typical Scottish weather, optical satellite retrieval is severely bottlenecked by cloud cover. Traditional products like LIS treat clouds as absolute barriers, resulting in a catastrophic drop in detection capabilities (Recall = 0.013) and a severe systematic underestimation of the snowpack (Snow-Only Bias = -37.23). While optical sensors inherently cannot penetrate opaque clouds, there are often views of snow through gaps in cloud, in cloud shadow, and in hazy conditions that can be utilised. By training SPUN on a lenient cloud mask (CloudScore+ < 0.5), we attempted to enable the model to maximise the retrieval of a useful signal. Because SPUN relies on learned contextual features rather than rigid spectral limits, it successfully identifies valid snowpack in poorly illuminated or hazy regions (Figure 5). Consequently, SPUN maintains a robust detection rate (Recall = 0.874, F1 = 0.747) and cuts the Snow-Only MAE in half compared to LIS (18.81% vs 37.78%). By accurately mapping snow right up to the cloud margins and within cloud shadows, SPUN extracts the maximum possible signal from the unoccluded portions of the scene. These results show promise that ML can learn complex non-linear decision functions that could resolve these long-standing issues with distinguishing snow and cloud in optical satellite imagery.

400 4.2 Generalisability, scalability, and a modular pipeline (Global Domain)

A primary concern with deep learning models trained on highly specific regional phenomena – such as late-lying Scottish snow patches – is catastrophic overfitting. To test this, we evaluated SPUN on the Wang et al. (2022) global dataset. We do note that within this dataset there are some cases of ice that is labelled as snow but not within the 6 test dataset images (see appendix figure C2). We have not tried to address the issue of snow and ice differentiation in this research.

405 Given these provisos, the SPUN pipeline achieves a performance of F1 Score 0.931, proving competitive with previous studies that were trained explicitly on this global dataset (Wang et al., 2022; Ma et al., 2023; Demil et al., 2025). This is a notable achievement; it indicates that the contextual representation of snow learned by SPUN in Scotland transfers successfully to the global context and the model is not overfitted to the training domain. To further evidence this point, we applied the SPUN model to the entire 40 images in the Wang et al. (2022) dataset and results were consistent achieving an F1 score of 0.900.

410 It is important to note the importance of cloud masking when transferring SPUN to the global context which we have outlined by the ablation study in Table 6. Without cloud masking the F1 score is reduced to 0.762 on the full dataset of 40 images, and this score reduction is attributable to the introduction of a significant number of false positives because of the cloud types in the global dataset that had not been seen within the Scotland training set and are similar to that spectral signature of snow. However, by decoupling the snow and cloud detection challenges, we enable SPUN to act as the "snow specialist" and CloudScore+ as the "cloud specialist" (but equally could be replaced by another cloud masking algorithm). This modularity allows users to define how strictly they want to mask clouds based on their specific need for precision versus sensitivity.



Gascoin et al. (2019) state, "[an operational product] should be scalable, i.e. allow the processing of large areas (10^4 km^2) with a reasonable computation cost (typically less than 1 h for a single product)". On one GPU, SPUN maps a 1 km^2 image patch in microseconds; in series, a whole S2 tile (12,100 image chips, that is $1.21 \times 10^4 \text{ km}^2$) takes ~ 4 minutes, but this is highly parallelisable so time could be significantly reduced. Consequently, even without exploiting parallel processing, SPUN exceeds this operational baseline by over an order of magnitude.

4.3 Training Data Challenges and Pseudo-Labelling

Creating an accurate and appropriately large, labelled dataset was a significant challenge for this study. Because manual labelling of high-resolution aerial imagery is time-consuming, we employed a "*pseudo-labelling*" strategy – training a separate U-Net model on cloud-free RGB aerial imagery to automatically generate additional masks. While this successfully expanded our dataset from 849 to 2,455 image chips (achieving a Dice score of 0.9578 on the aerial data), it resulted in a training set that was still fundamentally small for deep learning and given the focus on late-season patchy snow, the target snow class is fundamentally in the massive minority (as seen in Figure A). Furthermore, the pseudo-labelling model is not perfect and inevitably introduced label noise. The automated process occasionally misclassified bright rock (e.g., quartzite) as snow, and temporal mismatches occurred; in at least one instance, significant snowmelt took place in the two days between the aerial survey and the S2 overpass, leading to a direct conflict between the label and the satellite data.

To overcome these challenges of data scarcity and label noise, we deliberately compensated through specific model architecture choices. First, we selected a U-Net architecture, which is inherently well-suited for feature extraction from limited datasets. Second, we utilised a specially adapted hybrid loss function which combines both region-based (Dice) and pixel-based (MAE) metrics providing a regularisation effect that prevents the model from overfitting to label noise. Finally, we employed transfer learning via pre-trained foundation model weights; by starting with a network that already possesses a general feature representation of optical satellite imagery, we drastically reduced the volume of target data required. Ultimately, this architectural design outweighed the limitations of the dataset; as seen in Figure 5, SPUN effectively "outlearns" its noisy training set, overcoming pseudo-labelling errors to produce highly coherent predictions.

4.4 Future work

The ESA SnowCCI project and AlpSnow have been developing new operational algorithms for snow cover mapping, including a ML product ((Barella et al., 2022)) and a spectral unmixing based method (Keuris et al., 2023). In particular, spectral unmixing-based methods are powerful, and comparison to these is missing from this study. This omission is largely due to endmember requirements; while our Scottish snow spectra might perform well locally, they would likely not readily generalise to global snow using this technique.

The fact that we managed to create a model that generalises well to completely different regions using only a small, localised dataset from Scotland highlights the untapped potential of DL. It suggests that with a more representative, global dataset, these methods could help overcome persistent issues in optical snow cover mapping – namely, cloud and topographic shading, atmospheric disturbance, patchy, metamorphosed, and dirty snow. Furthermore, these techniques also show promise



for differentiating snow from ice, another long-standing issue (Aberle et al., 2025). A coordinated community effort to collate very high-resolution aerial and drone surveys of snow across diverse global environments would be invaluable. Because high-resolution masks can be aggregated to almost any satellite sensor (e.g., Planet at 5 m, S2 at 10 m, Landsat at 30 m), a single, robust, open-access dataset could provide the representative training data necessary to significantly reduce or overcome these challenges.

Finally, future work must address the practical barriers preventing widespread ML adoption within the broader snow mapping community. While ML is not a panacea, a significant gap remains between its capabilities and its current operational utilisation. A major hurdle is the encouraging open science practices; many ML studies in this domain do not publicly release their underlying code or training datasets, limiting reproducibility. Even when models are open-access, the technical barrier to entry remains steep for practitioners lacking specialised ML expertise. While developing fully operationalised software interfaces is often beyond the scope of individual research projects, the community must prioritise releasing transparent, well-documented code alongside accessible data pipelines. Furthermore, the "black box" nature of deep learning can hinder user trust and obscure the understanding of how physical processes are being interpreted. To bridge this disconnect, future ML initiatives should consider incorporating physics-informed learning and explainable AI techniques to demystify model behaviour for end-users, we have begun to explore this for SPUN in Appendix D.

5 Conclusions

This research successfully demonstrates that DL architectures offer a significant advancement over traditional spectral thresholding methods for optical snow cover mapping. To the best of our knowledge, we present SPUN as the first convolutional deep learning model designed to regress continuous percentage SCF directly from satellite imagery.

SPUN improves upon other snow mapping algorithms, including the operational Theia Let-It-Snow product, in the challenging, temperate maritime environment of the Scottish Highlands. Our model accurately captures patchy late-season snow and demonstrates remarkable operational resilience: where classical, rule-based spectral thresholds struggle with atmospheric distortion – often treating clouds and shadows as absolute barriers – SPUN successfully extracts valid hydrological signals. This robustness is achieved because the ML model learns complex, non-linear decision functions that leverage spatial and textural context, allowing it to identify snow through cloud gaps, atmospheric haze, and shadow.

Furthermore, the use of pre-trained foundation model weights enables the model to generalise effectively. When evaluated against a diverse global dataset, SPUN maintained highly competitive performance without catastrophic overfitting to the Scottish training domain, confirming that this methodology is not geographically limited to our specific study area.

Ultimately, SPUN provides a robust and scalable tool for small-scale studies of snow dynamics, hydrology, and ecology. More broadly, this research indicates that ML is a powerful, underutilised tool potentially capable of resolving persistent issues in operational snow mapping. Moving forward, the primary barrier to global application is data-driven; a coordinated community effort to collate diverse, high-resolution aerial surveys into a representative training dataset would allow future

<https://doi.org/10.5194/egusphere-2026-3163>

Preprint. Discussion started: 24 July 2026

© Author(s) 2026. CC BY 4.0 License.



models to fully exploit the available optical information, maximising snow detection in historically challenging, marginal, and cloud-prone environments.



Appendix A: SPUN Dataset Composition

Table A1. Detailed Dataset Composition. Absolute counts and relative proportions of images and pixels across the training, validation, and test splits. Percentages in parentheses represent the proportion of the respective total.

Split	Macro-Scale (Images)		Micro-Scale (Pixels)			
	Total	With Snow (%)	Total Pixels	Snow (%)	Cloud (%)	Background (%)
Train	1,052	312 (29.7%)	10,520,000	77,324 (0.74%)	1,083,344 (10.30%)	9,359,332 (88.97%)
Validation	469	39 (8.3%)	4,690,000	23,074 (0.49%)	1,549,664 (33.04%)	3,117,262 (66.47%)
Test	934	139 (14.9%)	9,340,000	32,424 (0.35%)	2,784,476 (29.81%)	6,523,100 (69.84%)



Two-Tiered Class Imbalance in SPUN Dataset

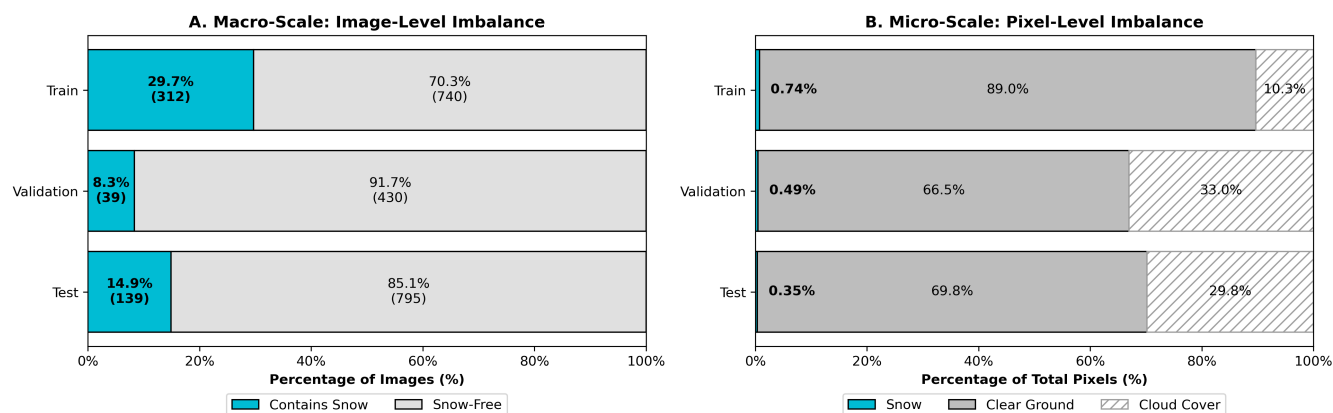


Figure A1. Two-tiered class imbalance in the SPUN dataset. (A) Macro-Scale (Image-Level Imbalance): Horizontal stacked bars display the percentage and absolute counts of S2 image chips containing any snow (shaded cyan) versus completely snow-free images (grey) across the train, validation, and test splits. (B) Micro-Scale (Pixel-Level Imbalance): Details the distribution of total pixels in each split into three classes: snow (cyan), clear ground (grey), and cloud cover (white with hatching pattern). Percentage labels inside the bars quantify the scarcity of snow pixels, ranging from 0.35% in the test set to 0.74% in the training data, relative to background terrain and prevalent cloud obscuration.



485 Appendix B: SPUN Model Architecture

Table B1. Ablation Study of Encoder Initialisations. Comparison of the Pareto-optimal trials for each pre-trained weight configuration evaluated during hyperparameter tuning. The selected configuration (DINO, Trial 15) is highlighted in bold, demonstrating the optimal balance between spatial segmentation accuracy (Dice) and continuous fractional error (RMSE) on the snow class. Arrows indicate the optimal direction for each metric.

Initialisation	Trial	Dice Score ($\uparrow$)	RMSE (Snow) ($\downarrow$)	Regression Loss Function	Learning Rate
DINO (Selected)	15	0.752	0.357	Huber	1.18×10^{-4}
ImageNet	19	0.750	0.361	MSE	3.90×10^{-5}
MoCo	73	0.748	0.375	MSE	2.54×10^{-4}
SoftCon	7	0.745	0.361	L1	8.28×10^{-5}
DeCUR	3	0.743	0.359	MSE	7.39×10^{-5}
Random	2	0.710	0.393	Huber	3.96×10^{-4}



Appendix C: Global Dataset

Geographical Distribution of Wang et al. (2022) Global Sentinel-2 Dataset

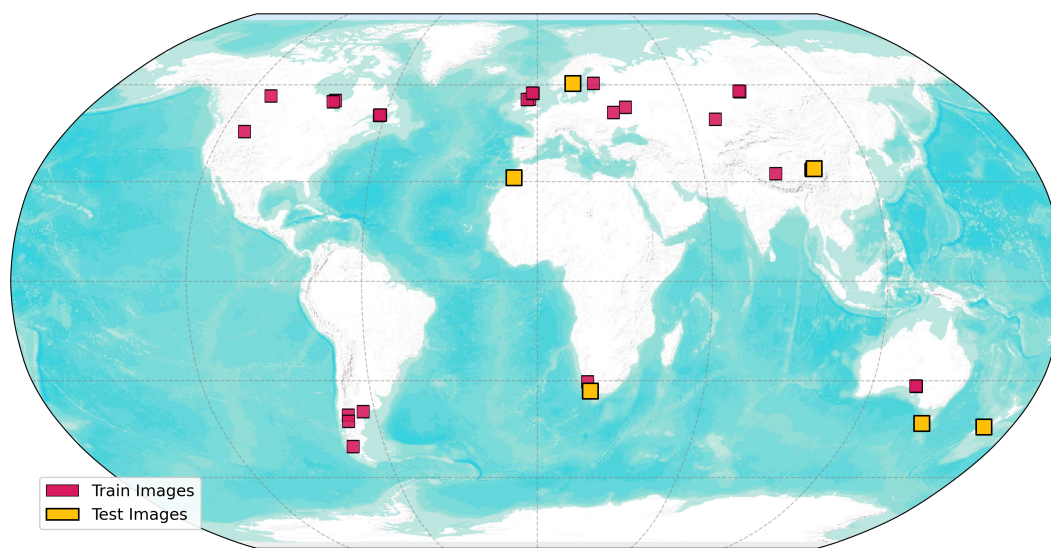


Figure C1. Geographical distribution of the 40 images within in the Wang et al., 2022 dataset. Their model was trained on the train images labelled (pink), and SPUN was compared to the performance of their model on the 6 test images labelled (yellow). SPUN was also assessed on the full dataset of 40 images. The ESRI World Terrain Base was using in creating this figure (https://server.arcgisonline.com/ArcGIS/rest/services/World_Terrain_Base/MapServer). Figure adapted from that found in Wang et al. (2022).

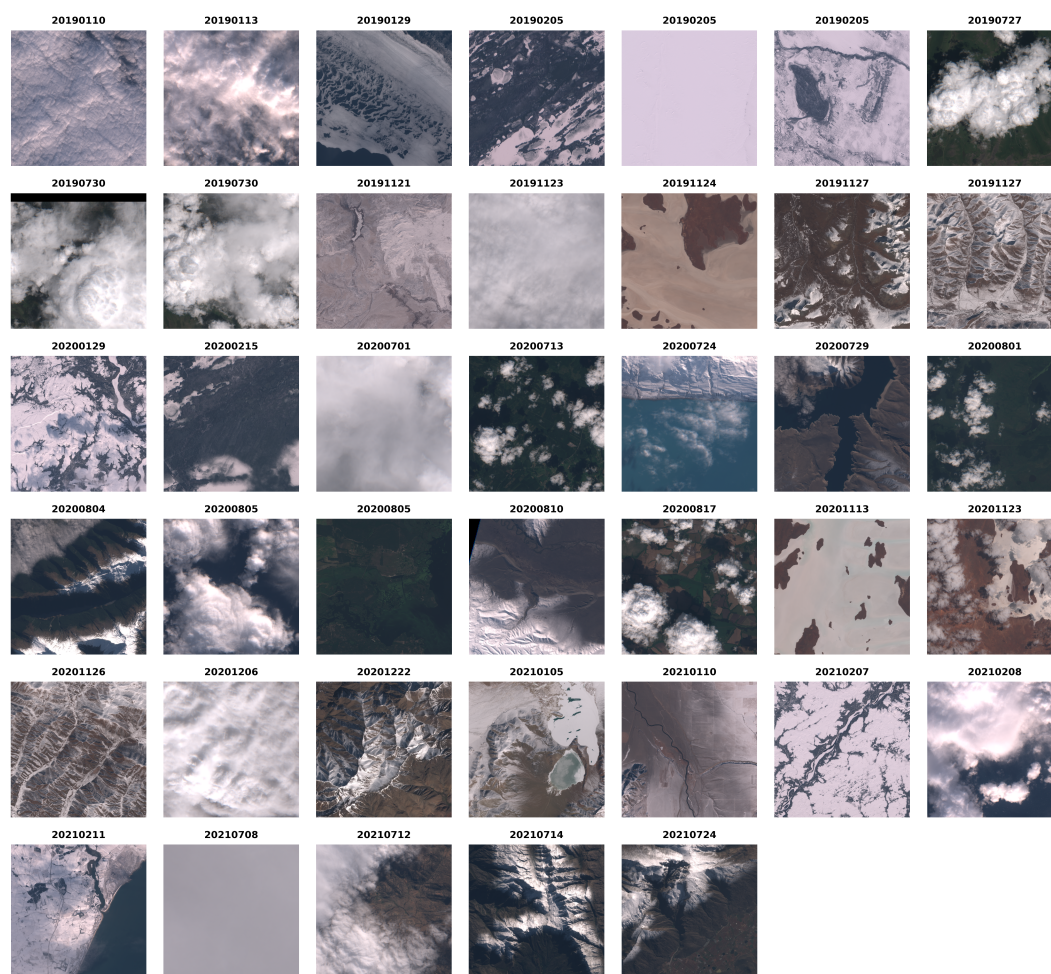


Figure C2. Sentinel-2 Level-1C data corresponding to the regions in the Wang et al. (2022) dataset. Figure adapted from that found in Wang et al. (2022).

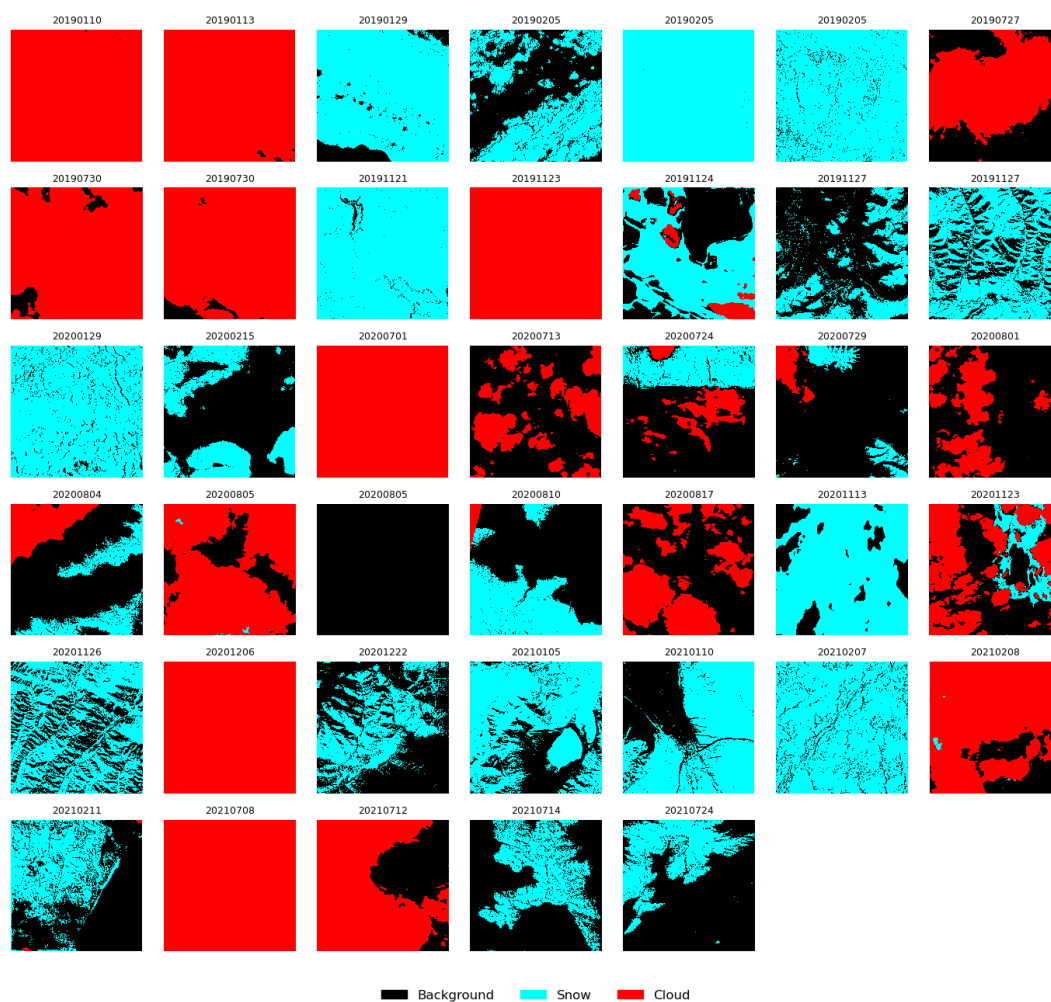


Figure C3. Binary snow maps produced using SPUN applied to all 40 images in the Wang et al., 2022 dataset. SPUN snow cover fraction maps are thresholded at 50% snow cover to create binary maps. Figure adapted from that found in Wang et al. (2022).



Table C1. Extended Robustness Assessment. Model performance metrics for the SPUN pipeline evaluated across all 40 images from the Wang et al. (2022) dataset. Metrics represent macro-averages across the Background, Cloud, and Snow classes. Baseline methods are excluded to prevent data leakage (Wang et al. (2022) utilised 34 of these images for training) and due to data unavailability (LIS).

Metric	SPUN (Ours)
Precision	0.913
Recall	0.918
F1-Score	0.915
mIoU	0.844

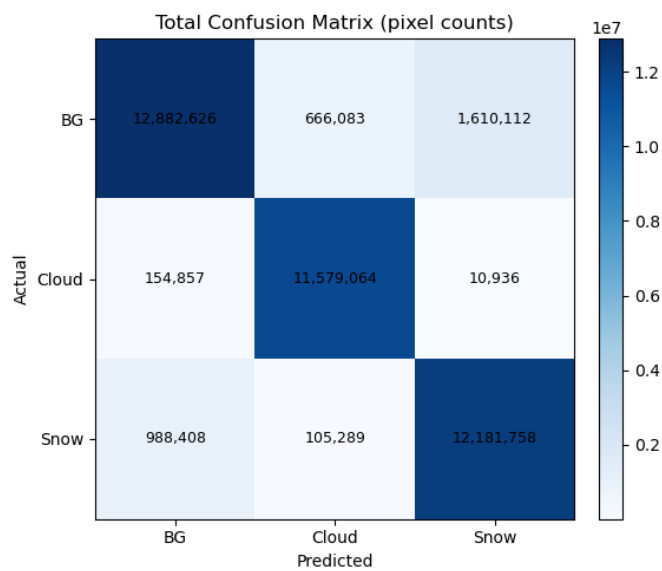


Figure C4. Confusion Matrix of SPUN performance on all 40 images from global dataset (Wang et al., 2022).



Table C2. Detailed Global Performance Comparison: Macro vs. Snow-Focused. This table compares the three methods evaluated natively in this study. The *Macro Avg* evaluates performance across all three classes (Background, Cloud, Snow). The *Snow-Only* metrics isolate the evaluation exclusively to pixels where either the prediction or the ground truth indicated snow. The bottom section ensures a direct image-to-image comparison by evaluating all three methods solely on the 4 test tiles where LIS operational data was available.

Method	Tiles (N)	Macro Avg (3-Class)				Snow-Only			
		Precision	Recall	F1	mIoU	Precision	Recall	F1	IoU
<i>Maximum Available Dataset per Method</i>									
SPUN (Ours)	6	0.941	0.924	0.931	0.874	0.956	0.837	0.892	0.806
Wang et al.	6	0.945	0.934	0.939	0.886	0.942	0.865	0.902	0.821
LIS (Ops)	4	0.830	0.814	0.803	0.672	0.912	0.771	0.835	0.717
<i>Direct 4-Tile Subset Comparison</i>									
SPUN (Ours)	4	0.933	0.923	0.927	0.866	0.918	0.840	0.877	0.781
Wang et al.	4	0.952	0.951	0.952	0.908	0.932	0.922	0.927	0.864
LIS (Ops)	4	0.830	0.814	0.803	0.672	0.912	0.771	0.835	0.717



Appendix D: SPUN Model Interpretability and Explainability

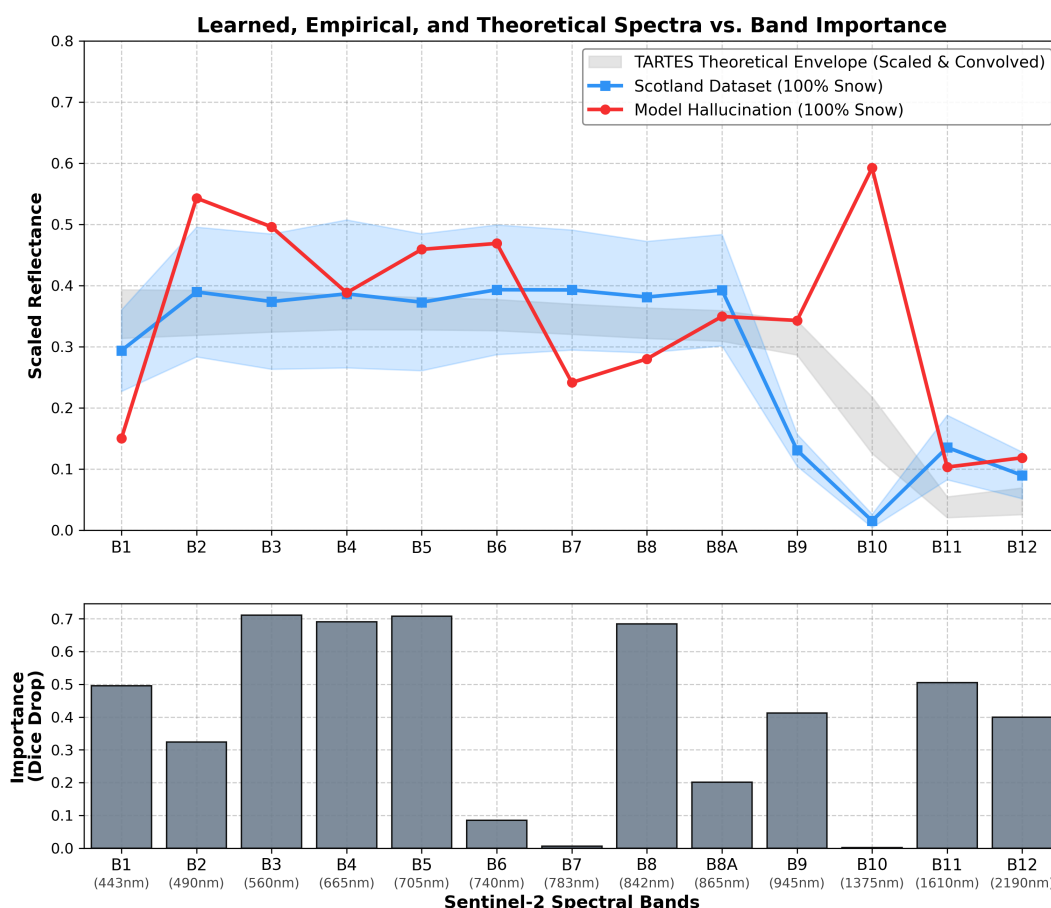


Figure D1. What is snow to SPUN? (Top) A comparison of SPUN’s internal representation of snow against empirical and theoretical baselines. The learned “Model Hallucination” (red) was generated via activation maximisation, optimising an input tensor to yield a 100% snow cover prediction. This implicitly learned spectrum largely aligns with the empirical mean of 100% snow pixels from the Scottish dataset (blue) and the theoretical reflectance envelope simulated using the TARTES radiative transfer model (grey, convolved with Sentinel-2 Spectral Response Functions). This alignment demonstrates that SPUN has accurately internalised the core radiometric signature of snow. Where the hallucination deviates from empirical reality (e.g. B10), it often occurs in bands the model deems unimportant (see Bottom panel); because SPUN learned to ignore these bands during training, their values are left unoptimised during activation maximisation. **(Bottom)** Permutation importance of the Sentinel-2 spectral bands, measured by the degradation in Dice score when each band is systematically noise-injected. The model correctly places weight on the highly reflective visible bands and the highly absorptive Shortwave Infrared (SWIR) bands (B11, B12), corroborating the physical necessity of these specific wavelengths for distinguishing snow from optically thick clouds. Furthermore, B8 is shown to be highly important; as the only Near-Infrared (NIR) band with a 10 m spatial resolution, this suggests the network leverages it specifically for resolving fine spatial boundaries.

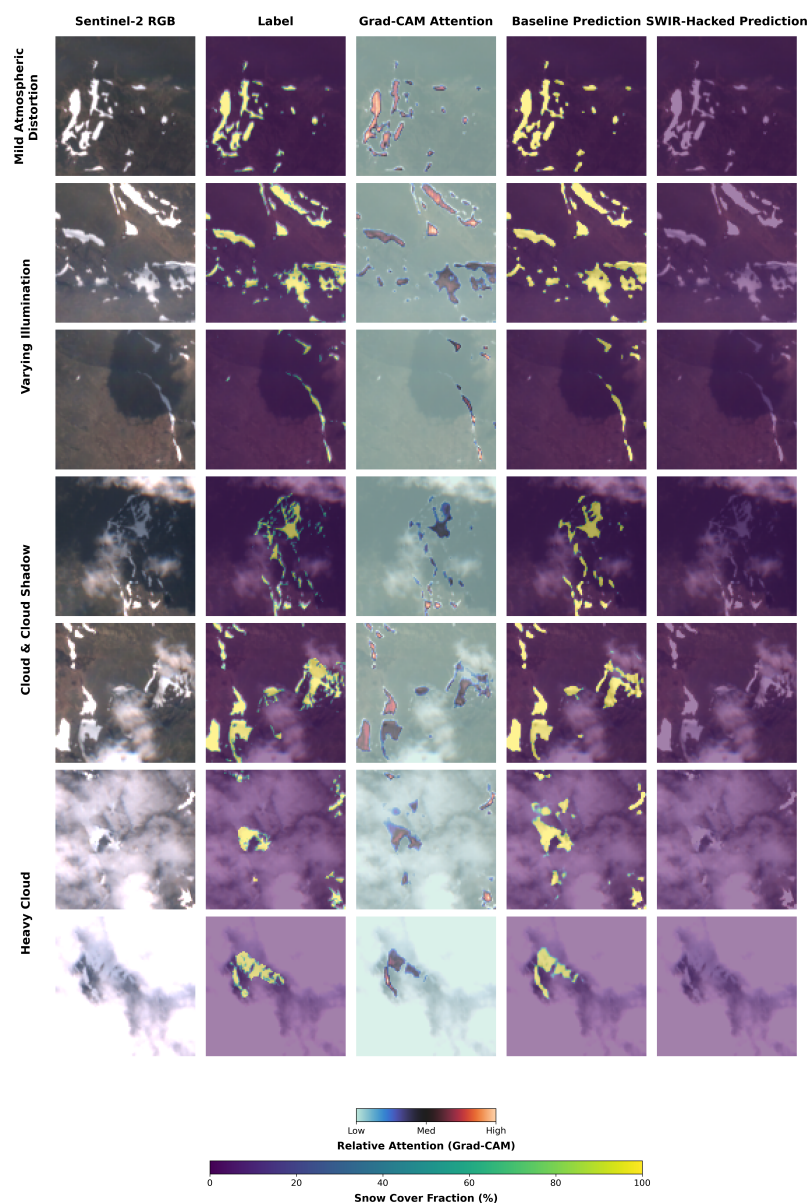


Figure D2. Spatial and physical interpretability of SPUN. (Columns 1 & 2) Sentinel-2 true-colour imagery and the corresponding labelled SCF map. (Column 3) Gradient-weighted Class Activation Mapping (Grad-CAM) visualises the spatial attention of the model. The attention mapping shows that higher attention is paid to the centre of snow patches, diminishing as pixels become more mixed towards the boundaries. Furthermore, attention decreases in topographic (rows 2 & 3) and cloud (rows 4 & 5) shade, reflecting the model's response to lower signal-to-noise ratios and Rayleigh blue-light scattering in these regions. (Columns 4 & 5) A counterfactual physical perturbation experiment. While the baseline prediction (Col 4) accurately estimates fractional snow cover, artificially saturating the Shortwave Infrared (SWIR) bands (B11, B12) to a reflectance of 0.8 to mimic an optically thick cloud causes the snow prediction to collapse to 0% (Col 5). This confirms that SPUN has internalised physical multispectral absorption rules (akin to the Normalised Difference Snow Index) to classify snow, rather than relying solely on visible band brightness.



Code and data availability. The code base for SPUN is available at <https://github.com/learnhowe/spun>. Release v1.0.0, which aligns with this manuscript, is archived on Zenodo at <https://doi.org/10.5281/zenodo.20495907>. The SPUN model weights and the full Scotland snow patch dataset can be found on Zenodo at <https://doi.org/10.5281/zenodo.20055817>.

Author contributions. LH developed the model, methodology, performed the analysis, and wrote the original draft. RE contributed to conceptualisation of the project and provided supervision and comments throughout. EJC provided comments on original draft. LH prepared the final manuscript with contributions from all co-authors.

Competing interests. The authors have no competing interests.

Financial support. This work was supported by NERC through a SENSE CDT studentship (NE/T00939X/1).

Acknowledgements. During the preparation of this manuscript, the authors used the AI tool Google Gemini to refine the text and improve readability. The authors thoroughly reviewed all AI-assisted edits and take full responsibility for the final published content.



References

- Aberle, R., Enderlin, E., O'Neel, S., Florentine, C., Sass, L., Dickson, A., Marshall, H.-P., and Flores, A.: Automated snow cover detection on mountain glaciers using spaceborne imagery and machine learning, *The Cryosphere*, 19, 1675–1693, <https://doi.org/10.5194/tc-19-1675-2025>, publisher: Copernicus GmbH, 2025.
- Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M.: Optuna: A Next-generation Hyperparameter Optimization Framework, <https://doi.org/10.48550/arXiv.1907.10902>, arXiv:1907.10902 [cs], 2019.
- Barella, R., Marin, C., Gianinetto, M., and Notarnicola, C.: A Novel Approach to High Resolution Snow Cover Fraction Retrieval in Mountainous Regions, in: *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, pp. 3856–3859, <https://doi.org/10.1109/IGARSS46834.2022.9884177>, iSSN: 2153-7003, 2022.
- Barnett, T. P., Adam, J. C., and Lettenmaier, D. P.: Potential impacts of a warming climate on water availability in snow-dominated regions, *Nature*, 438, 303–309, <https://doi.org/10.1038/nature04141>, publisher: Nature Publishing Group, 2005.
- Bavay, M., Lehning, M., Jonas, T., and Löwe, H.: Simulations of future snow cover and discharge in Alpine headwater catchments, *Hydrological Processes*, 23, 95–108, <https://doi.org/10.1002/hyp.7195>, eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/hyp.7195>, 2009.
- Beniston, M., Farinotti, D., Stoffel, M., Andreassen, L. M., Coppola, E., Eckert, N., Fantini, A., Giacona, F., Hauck, C., Huss, M., Huwald, H., Lehning, M., López-Moreno, J.-I., Magnusson, J., Marty, C., Morán-Tejeda, E., Morin, S., Naaim, M., Provenzale, A., Rabatel, A., Six, D., Stötter, J., Strasser, U., Terzago, S., and Vincent, C.: The European mountain cryosphere: a review of its current state, trends, and future challenges, *The Cryosphere*, 12, 759–794, <https://doi.org/10.5194/tc-12-759-2018>, publisher: Copernicus GmbH, 2018.
- Demil, G., Haghighi, A. T., Klöve, B., and Oussalah, M.: Seeing through the clouds: enhanced snow and cloud segmentation in sentinel-2 imagery with mDeepLabV3+, *Earth Science Informatics*, 18, 477, <https://doi.org/10.1007/s12145-025-01950-6>, 2025.
- Dozier, J.: Spectral signature of alpine snow cover from the landsat thematic mapper, *Remote Sensing of Environment*, 28, 9–22, [https://doi.org/10.1016/0034-4257\(89\)90101-6](https://doi.org/10.1016/0034-4257(89)90101-6), 1989.
- 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 Sensing of Environment*, 120, 25–36, <https://doi.org/10.1016/j.rse.2011.11.026>, 2012.
- Estilow, T. W., Young, A. H., and Robinson, D. A.: A long-term Northern Hemisphere snow cover extent data record for climate studies and monitoring, *Earth System Science Data*, 7, 137–142, <https://doi.org/10.5194/essd-7-137-2015>, publisher: Copernicus GmbH, 2015.
- Flanner, M. G., Shell, K. M., Barlage, M., Perovich, D. K., and Tschudi, M. A.: Radiative forcing and albedo feedback from the Northern Hemisphere cryosphere between 1979 and 2008, *Nature Geoscience*, 4, 151–155, <https://doi.org/10.1038/ngeo1062>, publisher: Nature Publishing Group, 2011.
- Gascoin, S., Grizonnet, M., Bouchet, M., Salgues, G., and Hagolle, O.: Theia Snow collection: high-resolution operational snow cover maps from Sentinel-2 and Landsat-8 data, *Earth System Science Data*, 11, 493–514, <https://doi.org/10.5194/essd-11-493-2019>, publisher: Copernicus GmbH, 2019.
- Gascoin, S., Luoju, K., Nagler, T., Lievens, H., Masiokas, M., Jonas, T., Zheng, Z., and De Rosnay, P.: Remote sensing of mountain snow from space: status and recommendations, *Frontiers in Earth Science*, 12, <https://doi.org/10.3389/feart.2024.1381323>, publisher: Frontiers, 2024.



- Hagolle, O., Huc, M., Desjardins, C., Auer, S., and Richter, R.: MAJA Algorithm Theoretical Basis Document, <https://doi.org/10.5281/zenodo.1209633>, publisher: Zenodo, 2017.
- Hall, D. K., Riggs, G. A., and Salomonson, V. V.: Development of methods for mapping global snow cover using moderate resolution imaging spectroradiometer data, *Remote Sensing of Environment*, 54, 127–140, [https://doi.org/10.1016/0034-4257\(95\)00137-P](https://doi.org/10.1016/0034-4257(95)00137-P), 1995.
- He, K., Zhang, X., Ren, S., and Sun, J.: Deep Residual Learning for Image Recognition, <https://doi.org/10.48550/arXiv.1512.03385>, arXiv:1512.03385 [cs], 2015.
- Hollstein, A., Segl, K., Guanter, L., Brell, M., and Enesco, M.: Ready-to-Use Methods for the Detection of Clouds, Cirrus, Snow, Shadow, Water and Clear Sky Pixels in Sentinel-2 MSI Images, *Remote Sensing*, 8, <https://doi.org/10.3390/rs8080666>, company: Multidisciplinary Digital Publishing Institute Distributor: Multidisciplinary Digital Publishing Institute Institution: Multidisciplinary Digital Publishing Institute Label: Multidisciplinary Digital Publishing Institute Publisher: publisher, 2016.
- Keuris, L., Hetzenecker, M., Nagler, T., Mölg, N., and Schwaizer, G.: An Adaptive Method for the Estimation of Snow-Covered Fraction with Error Propagation for Applications from Local to Global Scales, *Remote Sensing*, 15, <https://doi.org/10.3390/rs15051231>, company: Multidisciplinary Digital Publishing Institute Distributor: Multidisciplinary Digital Publishing Institute Institution: Multidisciplinary Digital Publishing Institute Label: Multidisciplinary Digital Publishing Institute Publisher: publisher, 2023.
- Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, <https://doi.org/10.48550/arXiv.1412.6980>, arXiv:1412.6980 [cs], 2017.
- Livneh, B. and Badger, A. M.: Drought less predictable under declining future snowpack, *Nature Climate Change*, 10, 452–458, <https://doi.org/10.1038/s41558-020-0754-8>, number: 5 Publisher: Nature Publishing Group, 2020.
- Lu, Y., James, T., Schillaci, C., and Lipani, A.: Snow detection in alpine regions with Convolutional Neural Networks: discriminating snow from cold clouds and water body, *GIScience & Remote Sensing*, 59, 1321–1343, <https://doi.org/10.1080/15481603.2022.2112391>, publisher: Taylor & Francis _eprint: <https://doi.org/10.1080/15481603.2022.2112391>, 2022.
- López-Moreno, J. I., Callow, N., McGowan, H., Webb, R., Schwartz, A., Bilish, S., Revuelto, J., Gascoin, S., Deschamps-Berger, C., and Alonso-González, E.: Marginal snowpacks: The basis for a global definition and existing research needs, *Earth-Science Reviews*, 252, 104 751, <https://doi.org/10.1016/j.earscirev.2024.104751>, 2024.
- Ma, J., Shen, H., Cai, Y., Zhang, T., Su, J., Chen, W.-H., and Li, J.: UCTNet with Dual-Flow Architecture: Snow Coverage Mapping with Sentinel-2 Satellite Imagery, *Remote Sensing*, 15, 4213, <https://doi.org/10.3390/rs15174213>, publisher: Multidisciplinary Digital Publishing Institute, 2023.
- Marsocci, V., Jia, Y., Bellier, G. L., Kerekes, D., Zeng, L., Hafner, S., Gerard, S., Brune, E., Yadav, R., Shibli, A., Fang, H., Ban, Y., Vergauwen, M., Audebert, N., and Nascetti, A.: PANGAEA: A Global and Inclusive Benchmark for Geospatial Foundation Models, <https://doi.org/10.48550/arXiv.2412.04204>, arXiv:2412.04204 [cs], 2025.
- Pasquarella, V. J., Brown, C. F., Czerwinski, W., and Rucklidge, W. J.: Comprehensive Quality Assessment of Optical Satellite Imagery Using Weakly Supervised Video Learning, pp. 2125–2135, https://openaccess.thecvf.com/content/CVPR2023W/EarthVision/html/Pasquarella_Comprehensive_Quality_Assessment_of_Optical_Satellite_Imagery_Using_Weakly_Supervised_CVPRW_2023_paper.html, 2023.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library, <https://doi.org/10.48550/arXiv.1912.01703>, arXiv:1912.01703 [cs], 2019.



- 570 Picard, G. and Libois, Q.: Simulation of snow albedo and solar irradiance profile with the Two-streAm Radiative TransfEr in Snow (TARTES) v2.0 model, *Geoscientific Model Development*, 17, 8927–8953, <https://doi.org/10.5194/gmd-17-8927-2024>, publisher: Copernicus GmbH, 2024.
- Ronneberger, O., Fischer, P., and Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, <https://doi.org/10.48550/arXiv.1505.04597>, arXiv:1505.04597 [cs], 2015.
- 575 Stewart, A. J., Robinson, C., Corley, I. A., Ortiz, A., Ferres, J. M. L., and Banerjee, A.: TorchGeo: deep learning with geospatial data, in: *Proceedings of the 30th International Conference on Advances in Geographic Information Systems, SIGSPATIAL '22*, pp. 1–12, Association for Computing Machinery, New York, NY, USA, <https://doi.org/10.1145/3557915.3560953>, 2022.
- Sturm, M., Holmgren, J., McFadden, J. P., Liston, G. E., Chapin, F. S., and Racine, C. H.: Snow–Shrub Interactions in Arctic Tundra: A Hypothesis with Climatic Implications, *Journal of Climate*, 14, 336–344, [https://doi.org/10.1175/1520-0442\(2001\)014<0336:SSIIAT>2.0.CO;2](https://doi.org/10.1175/1520-0442(2001)014<0336:SSIIAT>2.0.CO;2), publisher: American Meteorological Society Section: Journal of Climate, 2001.
- 580 Wang, Y., Su, J., Zhai, X., Meng, F., and Liu, C.: Snow Coverage Mapping by Learning from Sentinel-2 Satellite Multispectral Images via Machine Learning Algorithms, *Remote Sensing*, 14, 782, <https://doi.org/10.3390/rs14030782>, number: 3 Publisher: Multidisciplinary Digital Publishing Institute, 2022.
- Wang, Y., Braham, N. A. A., Xiong, Z., Liu, C., Albrecht, C. M., and Zhu, X. X.: SSL4EO-S12: A Large-Scale Multi-Modal, Multi-Temporal Dataset for Self-Supervised Learning in Earth Observation, <https://doi.org/10.48550/arXiv.2211.07044>, arXiv:2211.07044 [cs], 2023.
- 585 Wipf, S. and Rixen, C.: A review of snow manipulation experiments in Arctic and alpine tundra ecosystems, *Polar Research*, 29, 95–109, <https://doi.org/10.1111/j.1751-8369.2010.00153.x>, _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1751-8369.2010.00153.x>, 2010.