FROM SCOTTISH SNOW PATCHES TO GLOBAL SNOW COVER: A DEEP LEARNING APPROACH TO SNOW COVER FRACTION RETRIEVAL

Leam Howe*[1], Richard Essery[1], Elliot J. Crowley[2]

[1] *School of GeoSciences, University of Edinburgh, Edinburgh, UK*
[2] *School of Engineering, University of Edinburgh, Edinburgh, UK*

ABSTRACT: Optical snow mapping products are often developed and validated on regions with deep, clean, continuous snowpacks under clear skies. Maritime mountain ranges with marginal snowpacks suffer from particularly harsh conditions for optical snow observations: fragmented cover, spectrally degraded snow, and near-permanent heavy cloud influence. In the case of cloud, quality is usually assured with a conservative cloud mask, but this discards useful observations.

We quantify the cost of that trade-off and offer an alternative. Using 25 cm aerial surveys of the Scottish Highlands, we built an open snow cover fraction (SCF) reference dataset for Sentinel-2 comprising 2455 labelled 1 km chips from 37 scenes, deliberately retaining imagery through cloud gaps and heavy atmospheric interference rather than masking it away. On this dataset we trained SPUN, a U-Net with a ResNet50 encoder and a regression head that predicts continuous SCF per 10 m pixel.

Evaluated on a spatially independent test set at the 20 m operational grid, SPUN and the operational Let-It-Snow (LIS) product are separated modestly under *clear-sky* conditions (snow-only MAE 19.10 % versus 23.26 %; F1 0.606 versus 0.493). Under *all-sky* conditions the gap becomes distinct: SPUN improves slightly (MAE 18.81 %, F1 0.747), while LIS recall falls to 0.013 and it recovers 1.6 % of the reference snow-covered area. Aggregated to the 1 km chip scale, SPUN tracks snow-covered area with R^2 = 0.934 against -0.238 for LIS.

SPUN is trained only on late-season Scottish snow patches but, nonetheless, transferred to an independent 40-scene global dataset, reaching a macro-averaged F1 of 0.931 and a snow-only F1 of 0.892. We present this as a proof-of-concept rather than a finished operational tool, and argue that the limiting factor in optical snow retrieval is training data rather than model architecture.

KEYWORDS: snow cover mapping; fractional snow cover; remote sensing; Sentinel-2; deep learning; maritime snowpack; Scotland

1. INTRODUCTION

Snow cover extent is among the few snowpack properties observable at scale and at useful frequency. It constrains hydrological forecasts, provides the spatial validation data that distributed snow models otherwise lack, and in principle offers a synoptic view of where snow is lying across terrain that no observer network reaches. The Sentinel-2 constellation delivers 10 m imagery on a five-day repeat (Drusch et al. 2012), a resolution well matched to the sub-100 m variability that topography imposes on mountain snow (Gascoin et al. 2024).

Considerable effort has gone into snow mapping algorithms for Sentinel-2, yet persistent issues remain. Operational products still rely largely on threshold-based methods built around some variant of the Normalised Difference Snow Index (Hall et al. 1995), a design that is difficult to generalise across viewing conditions (Fig. 1). Four conditions are particularly challenging:

- **Mixed pixels.** As a snowpack fragments, an increasing proportion of pixels straddle a snow boundary. A binary snow/no-snow decision forces each of these to 0 or 100 %, discarding the sub-pixel information that matters most in patchy conditions.
- **Spectral degradation.** Dust, organic debris and grain coarsening move snow reflectance away from the assumed signature, particularly in the bands NDSI depends on.
- **Shadow and low sun.** Steep terrain and low solar elevation darken exactly the aspects that hold snow longest.
- **Cloud.** Snow and cloud are spectrally similar, so conservative masks are used to protect against misclassification — discarding partially usable scenes wholesale, including snow visible through gaps and haze.

Cloud deserves emphasis because it is usually treated as a data availability problem rather than an algorithmic one. Validation studies typically restrict themselves to clear-sky imagery, which systematically overstates operational performance (Gascoin et al. 2024). Development has

* *Corresponding author address:*
Leam Howe, School of GeoSciences, University of Edinburgh, Drummond Street, Edinburgh EH8 9XP, UK;
email: leamhowe1@gmail.com

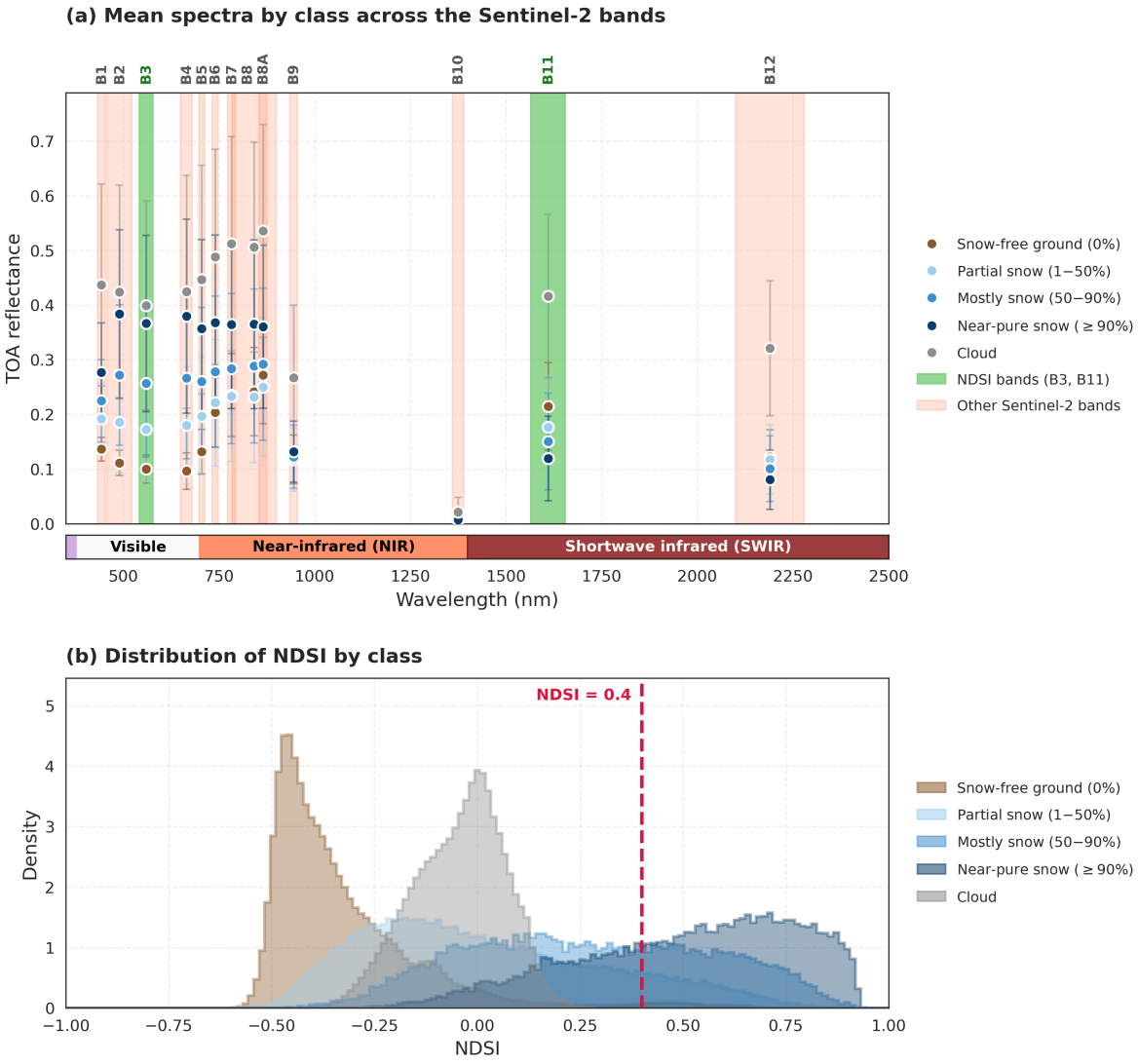

Figure 1: Sentinel-2 Level-1C top-of-atmosphere reflectance by class across all 13 bands (a), and the corresponding NDSI distributions (b), for the Scottish SCF dataset. Markers show class means with ± 1 standard deviation; the two bands used to compute NDSI (B3, B11) are highlighted in green. Of pixels with a reference SCF of 90 % or above, 33 % fall below the conventional 0.4 threshold; recomputing NDSI from Level-2A surface reflectance reduces this only to 30 %.

also concentrated on regions with persistent seasonal snowpacks, leaving marginal and ephemeral regimes under-represented (López-Moreno et al. 2024) even as warming pushes more regions toward them.

Machine learning (ML), and particularly deep learning, is advantageous for problems of this kind because it does not require pre-defined decision boundaries. Where a threshold must be set explicitly and applied uniformly, a learned model can infer a complex non-linear function directly from labelled examples, drawing on spatial context as well as spectral values. Here we present SPUN, a U-Net based model that uses semantic segmentation techniques from computer vision to regress SCF at the pixel level, and benchmark it against the operational LIS product (Gascoin et al. 2019).

We take the Scottish Highlands as a stress test, because Scotland presents all four challenging conditions at once: persistent Atlantic cloud, high-latitude sun angles, and a marginal snowpack that spends much of the season fragmented and heavily metamorphosed. It is an unrepresentative place to look for snow and an unusually informative place to test an algorithm. Our training data are late-season snow patches and are therefore not directly relevant to avalanche-season conditions, but the conditions those data forced the model to handle are not confined to Scotland. This paper condenses Howe et al. (2026), to which we refer readers for the full methods and analysis.

2. A FRACTIONAL SNOW COVER DATASET

ML-based retrieval requires labels, and fractional snow cover cannot be labelled by eye at satellite resolution. We therefore labelled at a resolution where the answer is unambiguous and aggregated upward.

The full labelling workflow is illustrated in figure 2. GetMapping distributes 25 cm aerial survey imagery of Scotland as 1 km tiles on the British National Grid. We manually labelled snow in a subset covering the Cairngorms, Ben Nevis and Glen Coe, yielding 849 image chips from 8 Sentinel-2 scenes, labelled in figure 3. To extend this without prohibitive manual effort, we trained a separate U-Net to segment snow in the 25 cm imagery (Dice 0.958, precision 0.986, recall 0.951 on held-out aerial data) and used it to generate *pseudo-labels* over further regions, expanding the dataset to 2455 chips from 37 Sentinel-2 scenes across 7 tiles and 7 years (figure 3).

Each labelled aerial tile was matched to the nearest Sentinel-2 acquisition within two days, and the 25 cm binary snow labels were resampled to 10 m fractional cover. Every satellite pixel therefore carries a snow fraction derived from roughly 1600 aerial pixels. We work with Level-1C top-of-atmosphere reflectance, both for comparability across sensors and because atmospheric correction is unreliable in the shadowed corries where late snow survives.

Cloud was identified using Google CloudScore+ (Pasquarella et al. 2023), applying the `cs_cdf` score at a lenient threshold of 0.5 rather than the more conservative values commonly used. This deliberately retains observations through cloud gaps and under heavy atmospheric interference so that the benchmark includes the marginal-visibility scenes on which snow mapping usually fails silently, and can quantify how much snow is recoverable in them.

The dataset was split into 1052 training, 469 validation and 934 test chips, as shown in figure 3. Splits are *spatially separate* and additionally separated by Sentinel-2 tile, so the test set represents unseen terrain rather than unseen dates over familiar ground.

3. THE SPUN MODEL

SPUN (Snow Patch U-Net) adapts the U-Net architecture (Ronneberger et al. 2015), in which a network progressively compresses an image to capture context and then expands it back to full resolution, with skip connections preserving fine spatial detail. The practical consequence is that classification draws on spatial context — the shape and setting of a patch — as well as the spectral values of a single pixel, which is what pixel-wise threshold methods are restricted to.

There are two key modifications to the standard U-net architecture. First, the usual classification output is replaced by a regression head predicting continuous SCF, making this to our knowledge the first application of deep learning to perform SCF regression from high-resolution (decametre-scale) optical satellite imagery. Second, training uses a combined Dice and regression loss: the Dice term rewards correct spatial extent, which matters when snow is a small minority of pixels, while the regression term penalises error in the fraction value itself.

The encoder is a ResNet50 initialised with DINO self-supervised weights from SSL4EO-S12 (Wang et al. 2023) — a satellite foundation model pre-

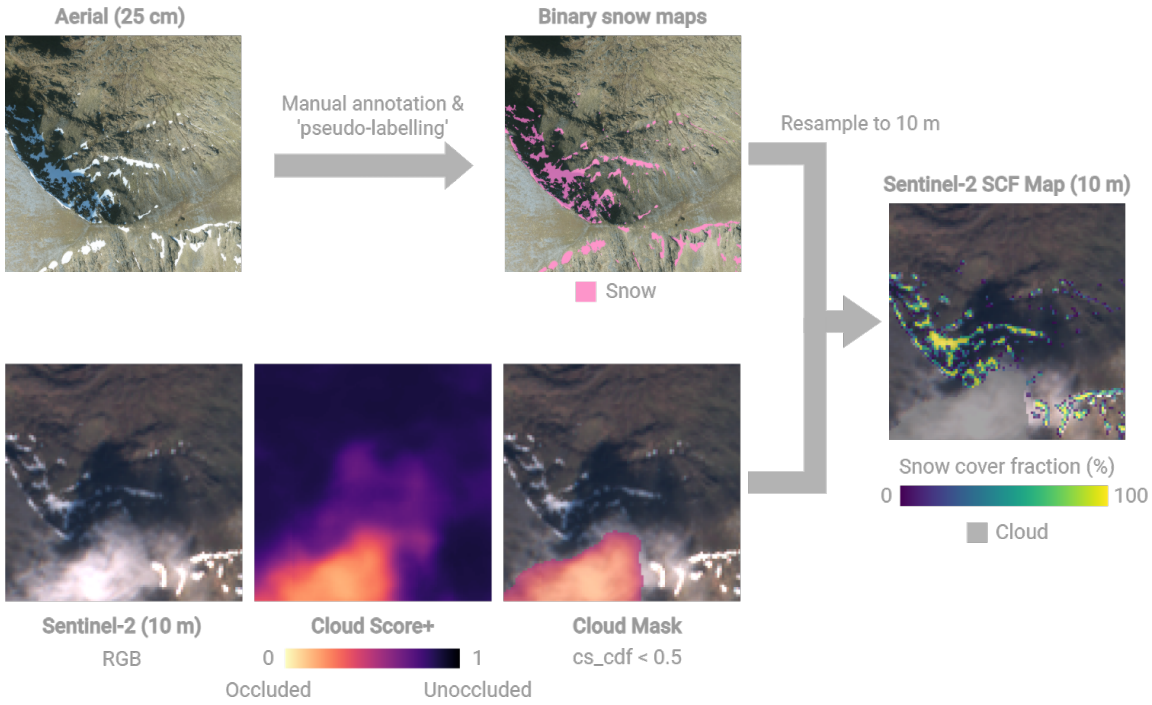

Figure 2: Labelling workflow to generate the SPUN Snow Cover Fraction (SCF) dataset. We attribute binary snow labels to 25 cm aerial data (© Getmapping Ltd) which is then resampled to generate 10 m SCF maps for Sentinel-2. Cloud is masked using the Google Cloud Score+ product thresholded at 0.5 to give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

trained on large volumes of unlabelled Sentinel-2 imagery. This gives the network a general representation of satellite scenes before it sees any snow labels, so less training data is required. Hyperparameters were tuned with Optuna over 100 trials (Akiba et al. 2019), selecting a Huber regression loss, Dice weight 0.166, learning rate 1.18×10^{-4} and batch size 16. Inputs are all 13 Sentinel-2 bands resampled to 10 m.

4. RESULTS

We evaluate at two levels. *Clear-sky* restricts the comparison to pixels the LIS cloud mask accepts, an example of the conditions under which snow products are conventionally validated. *All-sky* admits the full test set, so that the two together quantify how much snow information is lost to cloud masking rather than to misclassification. Comparison against the LIS product is made at its native 20 m grid.

4.1. *Clear skies: a modest advantage*

Under clear-sky conditions the two methods are separated but not dramatically (Table 1). SPUN reduces snow-only MAE from 23.26 % to 19.10 %, an 18 % relative reduction. The difference in character is more revealing than the difference in magnitude: LIS is the more conservative product, achieving slightly higher precision (0.562 versus 0.492) but recall of only 0.439, meaning it omits more than half the snow present under ideal viewing conditions. Its bias is strongly negative (-14.06 %). SPUN favours completeness, with recall of 0.789 and near-zero bias (+2.74 %).

This is a defensible design difference. A conservative product avoids false alarms, and for many applications that is the right trade.

4.2. *All skies: the operational reality*

Admitting the full test set separates the two methods entirely. SPUN's error is essentially unchanged (MAE 18.81 %) and its F1 improves to 0.747. LIS recall falls to 0.013 and its F1 to 0.026, because its cloud mask flags the large majority of the test data as invalid. Snow-only MAE rises to 37.78 % with a bias of -37.23 %.

Aggregated to the 1 km chip scale, SPUN tracks reference snow-covered area closely across the 165 snow-relevant test chips (R^2 = 0.934), recovering 118 % of it, while LIS recovers 1.6 % with no predictive skill (R^2 = -0.238). We emphasise that this is not a defect in LIS on its own terms. It is behaving as designed, declining to report snow it cannot see with confidence. The finding is about what

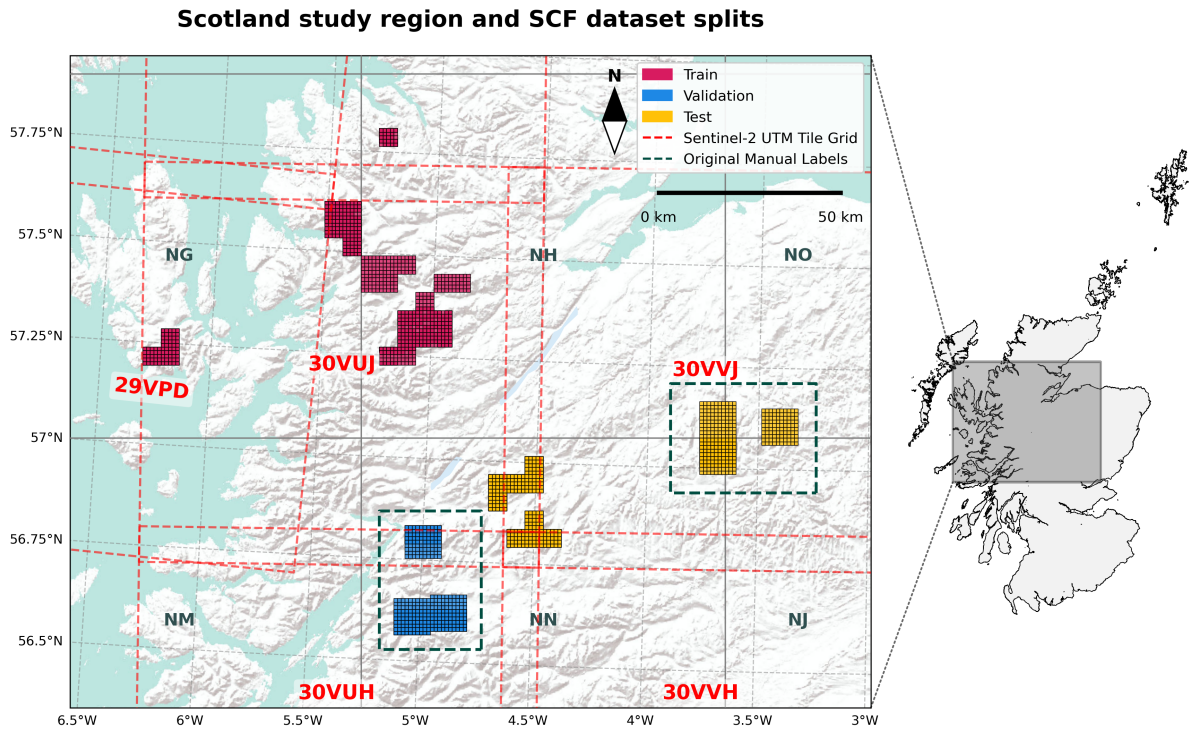

Figure 3: Map of study region and train/val/test data splits. These splits were done in an attempt to keep tiles from different S2 retrievals (red dashed lines) separate, whilst maintaining spatial separation and the balance of examples containing snow proportional within the train, validation and test 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.

that design costs under the viewing conditions this dataset samples: not a degraded answer, but effectively no answer, on most days. How much is lost over a larger region with more extensive snow is a different question, and one this test set was not built to answer — it deliberately concentrates on marginal views, where the proportion of snow lost to cloud masking is largest.

Figure 4 visualises the behaviour behind these numbers. SPUN retrieves snow through gaps in broken cloud, in cloud shadow, and in scenes with mild atmospheric distortion that LIS classifies as cloud outright.

4.3. *Transfer beyond the training domain*

Trained exclusively on late-season Scottish snow patches, SPUN might be expected to fail elsewhere. Applied without retraining to the independent 40-scene global dataset of Wang et al. (2022), spanning six continents, it achieved a macro-averaged F1 of 0.931 and mIoU of 0.874 across snow, cloud and background — at parity with the U-Net trained on that dataset directly (F1 0.939) and close to more recent transformer architectures (Ma et al. 2023; Demil et al. 2025). Snow-only F1 on the test subset was 0.892.

This result depends on the modular cloud mask. Without it, SPUN's false positive rate across all 40 scenes is 24.3 %, reflecting genuine confusion between global cloud types and those represented in Scottish training data; with CloudScore+ applied, the rate falls to 5.8 % while recall is preserved at 0.918. The honest reading is that SPUN is a strong snow detector paired with an off-the-shelf cloud mask, not a single model that has solved both problems.

Table 1: Performance on the spatially independent Scottish test set at the 20 m operational grid. MAE is snow-only, computed on the union of true and predicted snow pixels. F1, recall, and precision use a 50 % fractional threshold.

Condition	*Model*	*MAE (%)*	*F1*	*Recall*	*Precision*
Clear-sky	SPUN	**19.10**	**0.606**	**0.789**	0.492
Clear-sky	LIS	23.26	0.493	0.439	**0.562**
All-sky	SPUN	**18.81**	**0.747**	**0.874**	**0.652**
All-sky	LIS	37.78	0.026	0.013	0.562

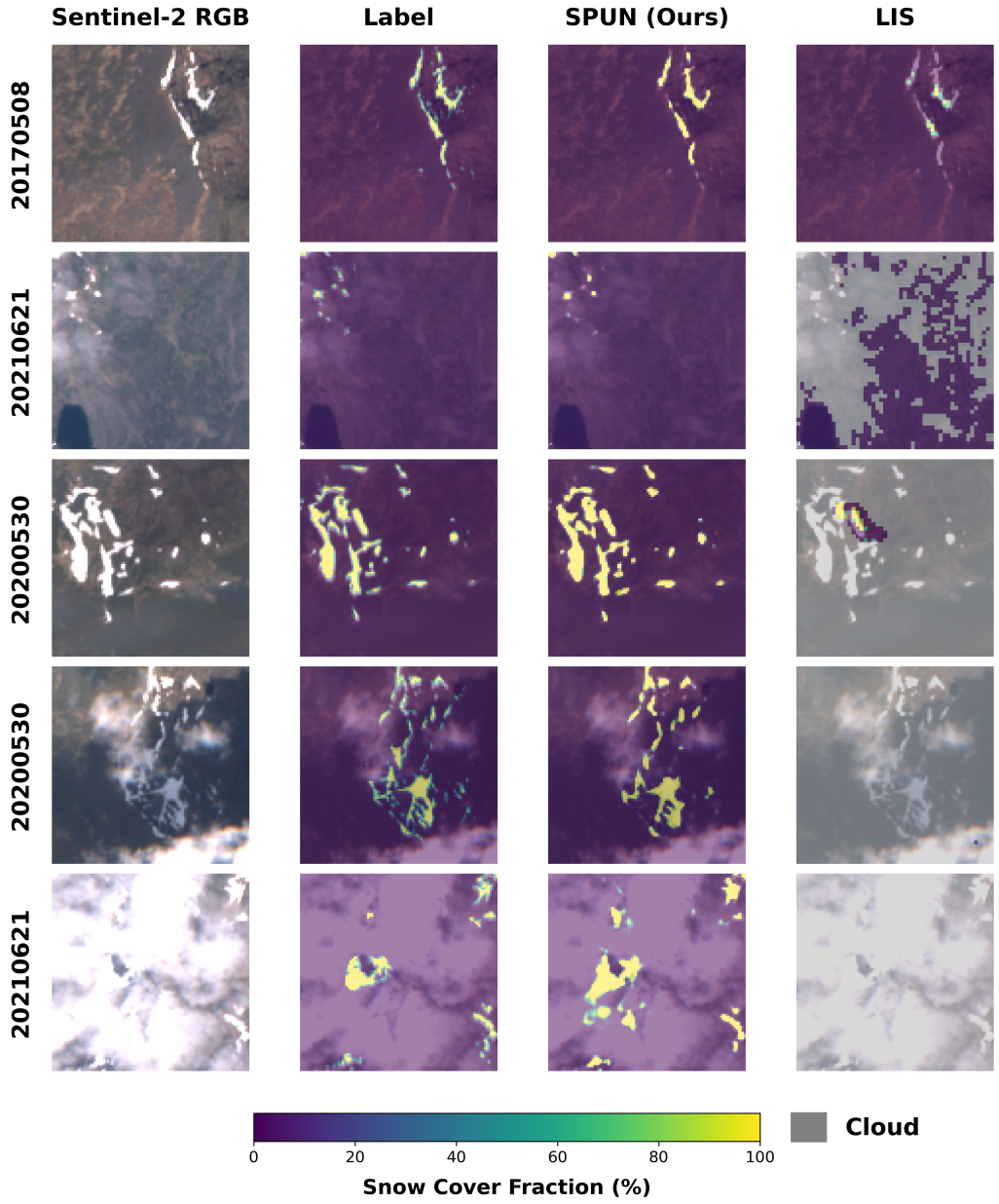

Figure 4: Visual examples of 1 km Sentinel-2 image chips from the test dataset. Each row shows a different image chip; each column shows a different snow mapping method. From left to right the columns are the Sentinel-2 RGB composite, the labelled data, the SPUN output, and the operational Let-It-Snow (LIS) algorithm. Row 1 is an example of well-illuminated snow patches, where algorithms are expected to perform well; row 2 shows small patches of snow with minor cloud influence and a lake; row 3 is an image with mild atmospheric distortion that LIS classifies as cloud; row 4 shows snow in cloud shadow; row 5 shows heavy cloud cover with views of snow through gaps.

5. LIMITATIONS AND PRACTICAL SCOPE

SPUN retrieves snow *cover*. It says nothing about depth, water equivalent, layering or stability, and it is not an avalanche forecasting tool.

Its training distribution is narrow, and failures align with where the dataset is lacking. Residual error is dominated by omission: on the global test subset, 16.3 % of true snow pixels are classified as background, against only 0.13 % of cloud pixels misclassified as snow. Boundary artefacts arise where strong topographic shadow falls across continuous snow cover. The model must then distinguish a shadow edge from a snow edge, and its training data contain few examples of the former: the aerial surveys were flown in spring and summer, so despite Scotland's latitude the dataset spans a narrow and favourable range of illumination geometries, and the snow it contains is patches on bare ground rather than continuous cover crossed by shadow.

These behaviours were identified by inspection of a decade of Sentinel-2 retrievals over the Cairngorms rather than quantified against reference data, since no fractional labels exist for those conditions. In our Cairngorms retrievals we therefore apply quality flags: scenes at low solar elevation are marked for manual checking, and scenes in which snow-covered area falls without corresponding warmth are flagged as physically implausible and likely to have lost snow to viewing conditions (Howe 2026).

Within its envelope — patchy, spectrally degraded snow under imperfect visibility — SPUN offers three practical gains over the operational product. It returns a usable snow map on far more days, which matters most where ground observations are sparse and cloud is the binding constraint on monitoring. It recovers more of the snow that is there: even under clear skies LIS omits over half of it, while SPUN recovers close to four-fifths. And it does so at 10 m rather than 20 m, four times the pixel density, which is the difference between resolving individual patches and averaging over them.

6. TOWARDS A COMMUNITY-DRIVEN GLOBAL SCF BENCHMARK DATASET

The clearest lesson from this work is that the ML architecture is not the limiting factor. That a model trained on a small, localised Scottish snow patch dataset generalises to completely different regions highlights the untapped potential of deep learning here, and where SPUN does fail (strong topographic shadow crossing continuous snow) the cause is a gap in the training data rather than the architecture. With more representative data, these methods could address persistent problems in optical snow mapping — cloud, topographic shading, atmospheric disturbance, and patchy, metamorphosed or dirty snow — and may extend to the long-standing difficulty of separating snow from ice (Aberle et al. 2025).

The reference data needed for this already exists, but is scattered and not always accessible: high-resolution snow maps are routinely produced for individual validation studies from Pléiades, WorldView, UAS and airborne lidar. A coordinated effort to collate and harmonise very high-resolution datasets across diverse environments would be invaluable, and because such masks can be aggregated to almost any sensor — Planet at 5 m, Sentinel-2 at 10 m, Landsat at 30 m — a single open dataset would serve all of them.

Our dataset, labels and code are openly available (Howe et al. 2026). We welcome discussion and feedback on this work and on the potential for bringing datasets like these together, and would equally like to hear of any efforts already underway.

ACKNOWLEDGEMENTS

This work was supported by the Natural Environment Research Council (grant NE/T00939X/1). Aerial imagery © Getmapping Ltd.

REFERENCES

Aberle, R., Enderlin, E., O'Neel, S., Florentine, C., Sass, L., Dickson, A., Marshall, H.-P., and Flores, A. (Apr. 2025). Automated snow cover detection on mountain glaciers using spaceborne imagery and machine learning. English. In: The Cryosphere 19.4. Publisher: Copernicus GmbH, pp. 1675–1693. ISSN: 1994-0416. https://doi.org/10.5194/tc-19-1675-2025.

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (July 2019). Optuna: A Next-generation Hyperparameter Optimization Framework. arXiv:1907.10902 [cs]. https://doi.org/10.48550/arXiv.1907.10902.

Demil, G., Haghighi, A. T., Klöve, B., and Oussalah, M. (July 2025). Seeing through the clouds: enhanced snow and cloud segmentation in sentinel-2 imagery with mDeepLabV3+. en. In: Earth Science Informatics 18.3, p. 477. ISSN: 1865-0481. https://doi.org/10.1007/s12145-025-01950-6.

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. (May 2012). Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. In: Remote Sensing of Environment. The Sentinel Missions - New

Opportunities for Science 120, pp. 25–36. ISSN: 0034-4257. https://doi.org/10.1016/j.rse.2011.11.026.

Gascoin, S., Grizonnet, M., Bouchet, M., Salgues, G., and Hagolle, O. (Apr. 2019). Theia Snow collection: high-resolution operational snow cover maps from Sentinel-2 and Landsat-8 data. English. In: Earth System Science Data 11.2. Publisher: Copernicus GmbH, pp. 493–514. ISSN: 1866-3508. https://doi.org/10.5194/essd-11-493-2019.

Gascoin, S., Luojus, K., Nagler, T., Lievens, H., Masiokas, M., Jonas, T., Zheng, Z., and De Rosnay, P. (May 2024). Remote sensing of mountain snow from space: status and recommendations. English. In: Frontiers in Earth Science 12. Publisher: Frontiers. ISSN: 2296-6463. https://doi.org/10.3389/feart.2024.1381323.

Hall, D. K., Riggs, G. A., and Salomonson, V. V. (Nov. 1995). Development of methods for mapping global snow cover using moderate resolution imaging spectroradiometer data. In: Remote Sensing of Environment 54.2, pp. 127–140. ISSN: 0034-4257. https://doi.org/10.1016/0034-4257(95)00137-P.

Howe, L. (Aug. 2026). https://github.com/leamhowe/spun. original-date: 2025-05-26T14:34:57Z.

Howe, L., Essery, R., and Crowley, E. J. (July 2026). SPUN: Deep learning for continuous snow cover fraction retrieval in marginal environments. English. In: EGUsphere. Publisher: Copernicus GmbH, pp. 1–40. https://doi.org/10.5194/egusphere-2026-3163.

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. (May 2024). Marginal snowpacks: The basis for a global definition and existing research needs. In: Earth-Science Reviews 252, p. 104751. ISSN: 0012-8252. https://doi.org/10.1016/j.earscirev.2024.104751.

Ma, J., Shen, H., Cai, Y., Zhang, T., Su, J., Chen, W.-H., and Li, J. (Jan. 2023). UCTNet with Dual-Flow Architecture: Snow Coverage Mapping with Sentinel-2 Satellite Imagery. en. In: Remote Sensing 15.17. Publisher: Multidisciplinary Digital Publishing Institute, p. 4213. ISSN: 2072-4292. https://doi.org/10.3390/rs15174213.

Pasquarella, V. J., Brown, C. F., Czerwinski, W., and Rucklidge, W. J. (2023). “Comprehensive Quality Assessment of Optical Satellite Imagery Using Weakly Supervised Video Learning”. en. In: pp. 2125–2135.

Ronneberger, O., Fischer, P., and Brox, T. (May 2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv:1505.04597 [cs]. https://doi.org/10.48550/arXiv.1505.04597.

Wang, Y., Braham, N. A. A., Xiong, Z., Liu, C., Albrecht, C. M., and Zhu, X. X. (May 2023). SSL4EO-S12: A Large-Scale Multi-Modal, Multi-Temporal Dataset for Self-Supervised Learning in Earth Observation. arXiv:2211.07044 [cs]. https://doi.org/10.48550/arXiv.2211.07044.

Wang, Y., Su, J., Zhai, X., Meng, F., and Liu, C. (Jan. 2022). Snow Coverage Mapping by Learning from Sentinel-2 Satellite Multispectral Images via Machine Learning Algorithms. en. In: Remote Sensing 14.3. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute, p. 782. ISSN: 2072-4292. https://doi.org/10.3390/rs14030782.