

Response to RC2

Leam Howe

October 13, 2026

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

Authors: Leam Howe, Richard Essery, and Elliot J. Crowley

Journal: Hydrology and Earth System Sciences (HESS)

We thank Simon Gascoin for their constructive review of our manuscript. Our point-by-point response is provided below with the original reviewer comment is copied in *italic* and our response to each point.

Point-by-point response

Major comments

This study is thoroughly conducted, and I have no significant concerns (except the feeling that my own work is on its way to becoming a digital fossil). The authors effectively addressed the challenge of obtaining sufficient training data by employing a smart pseudo-labelling methodology from aerial imagery. As noted by the authors (L451), this approach has the potential to be scaled up on a larger scale through a coordinated effort to segment aerial or VHR satellite imagery from national agencies (e.g. a European COST action!). However, the main limitation of this study is also related to this training dataset of late-snow season images in Scotland. The generalization of the model was tested using a limited set of products from different regions across the world, but the discussion focused on the geographical diversity, not seasonality. Early winter snow detection is notoriously difficult especially in high latitude regions due to the low solar elevation. I wonder if the authors could discuss the performance of the GS+/SPUN pipeline in these conditions? They could argue that it is out of the scope of their study since they clearly target end-of-season patchy snow in their abstract and introduction. However, a significant part of the paper shows that SPUN can do much more than that. This leads me to my second main comment: Ultimately what is the goal of SPUN? Is it to become the next-generation tool for operational snow mapping? Or is it a methodological step to characterize late-season snow in Scotland as accurately as possible? If the latter is the case, the significance of late-season snow in Scotland and the perspective of the study could be more thoroughly explained. The introduction is somewhat vague in this regard. I doubt that snow patches in Scotland substantially “influence the global energy budget”. However, it is well-established that late-season snow covered area distribution reflects winter accumulation, making it a valuable information for reconstructing SWE and thereby enhancing ecohydrological models in this region.

Author response: Thank you for the kind words and thoughtful insight. I do believe that a coordinated effort to collate and label VHR imagery would be invaluable, and a European COST action would be an interesting avenue to discuss further.

We agree that seasonality is not thoroughly explored in this study and would be hugely beneficial. We argue that further investigation of this is outside the scope of this study but an important question for future research and scaling up of similar machine learning

approaches. However, we have qualitatively investigated the SPUN/CS+ pipeline on a decade of data for the Cairngorms National Park, Scotland, and note that performance is impacted by these ‘notoriously’ difficult viewing conditions. Specifically, the SPUN pipeline struggles with harsh shadows over extensive snow cover; the sharp contrasting boundaries, we believe mess with the convolutional mechanism with which SPUNs predictions rely on, leading to boundary artifacts. Quantifying this effect is beyond the scope of this study, but we have created a dashboard to view our dataset of snow maps for the Cairngorms which are linked to through the SPUN GitHub repo.

As for the ‘ultimate goal of SPUN’, it is a tool intended for mapping late-season patchy snow, and was initially intended for the author’s research to better quantify late season snow in Scotland. However, we believe that the generalisability shown by this model indicates the untapped potential of machine learning methods for larger scale, and indeed perhaps for next generation operational snow mapping, and wanted to acknowledge that alongside presenting the model as is. We agree with the reviewer here that we have lost the message of the intended scientific motivation behind SPUN in the first place and we will improve the introduction to better reflect this.

Minor comments

L77 “challenging” indeed, but it would be even more challenging if dense evergreen forests were present. The authors should describe the vegetation in their study area, maybe discuss this as a potential limitation?

Author response: Agreed, there is limited-to-no tree cover in our study region and this is an important limitation which needs to be acknowledged. We will add this to the text.

L96, 122, 192: specify the resampling method

Author response: We apologise for overlooking these details and agree that including them is important for reproducibility. We will add this, see response to RC1 for more details.

128: shown by who?

Author response: This is information provided in an article provided alongside the product on the Earth Engine Data Catalog. We will cite the article and dataset in the manuscript.

145: The performances are already excellent, but I wonder if the authors considered adding topographic data in input tensors. I also wonder if the model “knows” image metadata such as the time of acquisition and the geographic coordinates? A DEM and solar angles could enhance the ability of the model to account for the effect of terrain slope and acquisition time on the spectral signature of the snow cover (see also the comment by Carlo Marin about Figure 1). A shaded DEM could help reduce false positive due to the spectral confusion between clouds and snow (because a cloud does not look like the hillshade). This may not be relevant for isolated snow patches though.

Author response: We did consider adding a DEM as a layer in the input, but believed that this information would cause overfitting to the regional climate. For example, we are training on Scotland where the snow lasts longest in the NE facing corries; you are therefore encoding this information into the model and learning where snow lasts in Scotland, not what snow looks like in Sentinel-2 data. There are definitely possibilities to get around/limit this overfitting including data augmentation (e.g. random flips and rotations) which we do already implement for SPUN’s training, so maybe it wouldn’t be a significant problem. We acknowledge that this is an avenue worth exploring, especially given the availability and improving resolution of global elevation data; however, believe this to be beyond the scope of this study. Modelled hillshade based on the theoretical viewing conditions is especially an interesting thought that we didn’t consider and could potentially help ‘learn’ what snow should look like at different times of year with different solar illumination angles – an exciting avenue for future research! Thank you for the thoughtful insight.

L389: not exactly. LIS tries to recover some pixels underneath MAJA’s cloud mask, but under very strict criteria.

Author response: Apologies for the incorrect description of LIS handling. We will update this to correctly reflect the cloud handling mechanism of LIS.

L407: it would be interesting to discuss the seasonal transferability as well if the Wang dataset allows.

Author response: An interesting point, but we argue this is beyond the scope of this study.

L419: Does this time also include model weights loading, data I/O, GC+ preprocessing, patch tiling, and mosaicking back into a full tile product? Also, GPUs are less common on cloud EO platforms, it would be interesting to also report inference time on CPU.

Author response: Yes, and retrieval of desired tile from GEE. A fair point about CPU vs GPU performance, we will check the time difference when run on CPU and report the values. Expect slower but still operationally viable.

Appendix D: I understand that the 100% B10 reflectance has no physical meaning, as this band does not provide information on the state of the land surface. Yet, the high reflectance of B9 is surprising to me. I suspect it is due to high humidity in the atmosphere above Scotland that was seen by the model in the training dataset. However, this may lead to omissions errors in arid regions, but also in high elevation regions where the atmosphere is thin. I wonder if this could be spotted while testing SPUN outputs against the Wang dataset (maybe I missed it).

Author response: We thank the reviewer for this observation, which prompted two additional analyses.

On the hallucinated spectrum itself, we should clarify how it is generated. Activation maximisation recovers the direction in which each band increases the predicted fraction, not a physically realisable spectrum. B10 makes this concrete: it attains the highest value of any band in the figure, yet its permutation importance is 0.002 and perturbing it changes F1 by less than 0.001. The model has effectively learned to ignore B10, which is precisely why the optimiser is free to assign it an arbitrary value. We repeated the optimisation from eight random initialisations and the spectrum is reproducible ($CV \leq 0.05$ in all bands), so these values are stable, but they are not the model’s estimate of snow reflectance. We will revise the caption to state this explicitly and will remove the claim that the agreement with the empirical spectrum demonstrates SPUN has internalised the radiometric signature of snow.

On the underlying concern, the reviewer is right that SPUN’s calibration to B9 is tied to the Scottish atmosphere, though the symptom differs from the one anticipated. We scaled each band multiplicatively across the test set, spanning the range between the observed Scottish B9 level and the unattenuated surface value implied by our TARTES simulations. Raising B9, as a drier or thinner atmosphere would, left recall over labelled snow unchanged (+0.007 at $\times 1.5$) but caused a sharp loss of precision beyond approximately $\times 1.2$, through false positives over bare ground. The error mode is therefore commission rather than omission. Brightening B3 and B5 produces qualitatively similar behaviour and brightening B11 and B12 the opposite, improving precision, consistent with the expected SWIR contrast between snow and other bright surfaces.

We agree there is potential to further investigate atmospheric transferability with the Wang dataset but argue that this is beyond the scope of this study.