Reviewer 1:

- The manuscript presents research on automated detection of mass movement-related phase patterns in Sentinel-1 wrapped interferograms using deep learning (DL) approaches. The authors established a custom-labelled training dataset containing over 1,000 manually mapped coherent phase patterns from four interferograms covering the canton of Valais, Switzerland, and benchmarked multiple semantic segmentation architectures. A domain-expert variability study was also conducted across ten selected case studies to quantify task-intrinsic uncertainty and contextualize DL performance. The work is methodologically thorough and of clear relevance to the landslide remote sensing community. However, several aspects of the experimental design and discussion require improvement before the manuscript can be accepted for publication. The following comments are offered to help strengthen the work:

We thank the reviewer for their valuable time and effort. We appreciate how the comments are thorough, solid and constructive. In the following, we respond to each comment individually.

- The guiding instruction provided to participants, *"Delineate the mass movement-related phase patterns that you would expect an automated mapping approach to detect on this interferogram"*, may unintentionally conflate two distinct cognitive tasks: geoscientific interpretation of the interferogram and subjective expectation of DL model capability. Because experts likely differ substantially in their assumptions about what automated systems can or cannot detect, a portion of the observed low inter-expert IoU (ranging from 0.18 to 0.49) may reflect inconsistency in perceived model capability rather than genuine disagreement in geomorphological judgment. This distinction is important, as it affects how the expert variability results should be interpreted and used as a benchmark for DL performance. The authors are therefore encouraged to explicitly discuss these two potential sources of disagreement in Section 4.1, and to clarify what design choices were made during the study setup to minimize this ambiguity in participant annotations.

We thank the reviewer for raising this important point. We understand the concern that the phrasing “that you would expect an automated mapping approach to detect on this interferogram” could be interpreted as conflating two distinct cognitive tasks:

1. geomorphological interpretation of visible phase patterns and
2. a subjective assessment of what a DL model might be capable of detecting.

In practice, the instructions provided to participants were more specific than the manuscript currently conveys. Participants were asked to map visible mass movement–related phase patterns, delineate the landslide boundary according to the observable signal, and base their delineation solely on the interferogram itself, without relying on external information or assumptions about landslide morphology. The intention behind the phrasing referring to an “automated mapping approach” was therefore not to invoke subjective expectations about model capability, but rather to emphasise that annotations should be derived strictly from the observable

phase signal in the interferogram. This constraint was introduced because most of the existing landslide inventories delineate landslides using additional information rather than strictly the interferometric phase signal, which would diverge from the information available to the DL model. The goal of the instruction was therefore to align the expert annotations with the input data provided to the model.

We agree that this intention is not sufficiently clear from the current wording in the manuscript. In the revised version we will clarify the instruction provided to participants and rephrase it as follows: “Delineate all mass movement–related phase patterns visible on this interferogram in the area of interest, without using other sources of information (such as orthoimagery or DEMs), and provide a fringe count where you feel confident to do so.”

- As noted in Section 2.3.2, all training labels were produced by a single individual, and the multi-round quality control described in Section 4.2.1 was also performed exclusively by the same mapper. Given that pairwise IoU values among the six experts ranged from as low as 0.18 to 0.49, this raises a substantive concern that the DL models may have learned the idiosyncratic annotation style of one operator rather than generalizable geoscientific characteristics of mass movement phase patterns. While the authors acknowledge this limitation in Section 4.2.1, no concrete mitigation strategy is presented. The manuscript would be considerably strengthened if the authors either: (a) engaged a second annotator to independently label at least a representative subset of the training data and reported cross-annotator consistency using IoU or a comparable metric; or (b) conducted a repeat-labelling exercise by the primary mapper on a subset of scenes to quantify intra-rater reliability. Either approach would provide a more transparent characterization of the label quality and its potential influence on model behavior and generalizability.

We thank the reviewer for this important comment and agree that the use of a single annotator introduces operator-specific bias. It is with this thought that we started with the expert analysis to try to quantify this subjectivity. It was, however, asking too much from any one of the invited experts to perform extensive mapping which is why we restricted ourselves to the selected case studies.

During label generation we implemented a multi-round quality control procedure in which all annotations were revisited multiple times from varying starting locations within the interferograms thereby reducing systematic bias due to e.g. fatigue effects. We also used two independent cyclic colourbars during the mapping process (“roma0” and “phase” (as implemented in QGIS)) to minimise bias due to colour-distinctability further reducing biases. As a first step, we will make all these precautions more clear in the revised manuscript.

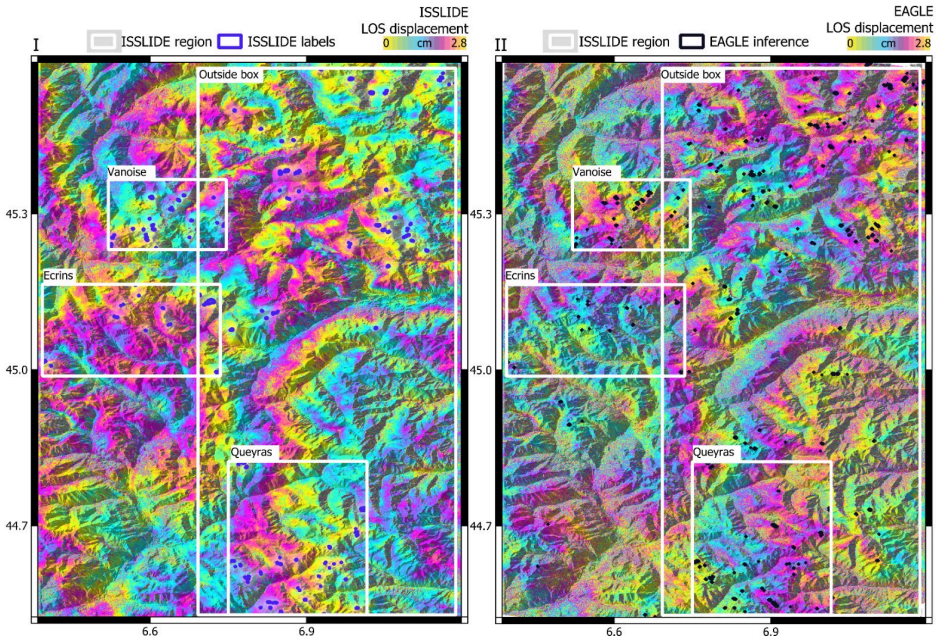
We acknowledge the reviewer’s concern however that the resulting bias and applicability for others could be better quantified. However, increasing the number of annotators would not solve the problem, as labeling in this specific context (mass movements outlines on InSAR phase patterns) can always be identified as arbitrary. For the revised manuscript, we propose to add inference results of our model on the region of the ISSLIDE database (Bralet et al., 2024), which we already reference in our manuscript, and compare to the manual labels provided in the

ISSLIDE database by independent annotators. With these new results, we provide a quantification of our model performance not only on our labels using cross-evaluation but also when compared to an external manually labelled dataset, in a different geological area that our model has not seen in the training. This way readers will get a better sense of the applicability, independently of the manual labeller. Since the ISSLIDE database does not offer exact mapping-outlines we can only report Recall (true positive rate) metrics.

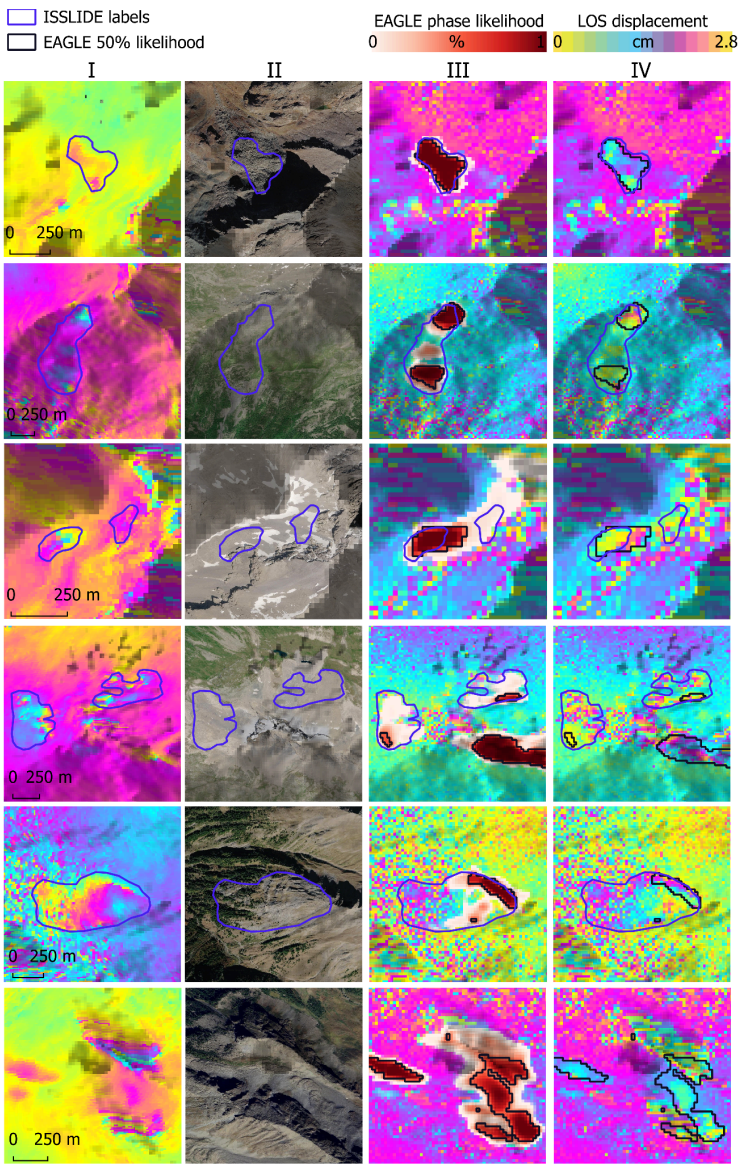
We will supply this information in subchapter 3.3: “To extend the comparison beyond the variability-study, we further applied the inference pipeline to the area covered by the ISSLIDE dataset (Bralet et al., 2024). The model recovered 41.85% of the manually labelled pixels (pixel-wise recall at 50% likelihood) and correctly detected 49, 39 and 24 of the 88 reference polygons, when setting an object-detection recall threshold at 25%, 50% and 75%, respectively."

The details and a visual of this will be added in the appendix – to also show that 1) while we cannot provide metrics on false-positives, the model is providing meaningful results 2) while we optimised to having the closest-as-possible processing pipeline as ISSLIDE for the interferogram generation, they are still different and so there was additional uncertainty introduced by using a different processing chain.

New Appendix H figure:



Appendix Fig. H.A: I. ISSLIDE interferogram (01.08.2018 – 13.08.2018) with manual labels (Bralet et al., 2024) and shapefile extents in comparison to II. our interferogram (also 01.08.2018 – 13.08.2018) and detected mass movements using the inference pipeline. Runtime of inference: 1min 20s. Image contains modified Copernicus Sentinel-1 imagery and hillshade derived from the Shuttle Radar Topography Mission (dataset SRTMGL1) and is displayed in WGS84.



Appendix Fig. H.B:Zoom into examples I. Zoom into ISSLIDE interferogram shown in previous image) for different examples showcasing the manual label, the II. corresponding scene as an optical image with the ISSLIDE label as reference III. the interferogram we processed and the inference outcome as both likelihood raster (mean over the five trained models from cross-evaluation) and 50% likelihood polygon in comparison to ISSLIDE label and IV. our interferogram with ISSLIDE label and 50% likelihood from our inference. Image contains modified Copernicus Sentinel-1 imagery, information derived from Shuttle Radar Topography Mission (dataset SRTMGL1) and Google Earth. Maps are displayed in WGS84 projection.

- The decision to discard all negative patches during training (Section 2.3.4), motivated by concerns about model conservatism, warrants further scrutiny in light of the precision results reported in Section 3.2.1. At an IoU threshold of t = 0.4, the model achieves a precision of approximately 64%, implying that roughly 36% of predicted detections do not correspond to true mass movement signals. By withholding negative examples

entirely during training, the model may have had limited exposure to confounding signals, such as atmospheric artefacts, local noise patterns, or geometric distortions, that can visually resemble coherent phase patterns. Furthermore, since the training labels were produced by a single mapper and the magnitude-frequency analysis demonstrates systematic undersampling of smaller mass movements below approximately 5,600 m² (Section 2.3.2, Fig. 5), patches discarded as negative due to the absence of annotations may in fact contain unmapped or sub-threshold deformation signals, introducing a subtle but non-trivial form of label contamination into the training process. This issue is well recognized in the broader landslide prediction literature, where the definition and selection of true negative samples has been shown to substantially affect model behavior and generalizability (e.g., https://doi.org/10.1007/s10346-020-01473-9; https://doi.org/10.3390/rs15123200). The authors are therefore encouraged to investigate this potential trade-off more rigorously, either by incorporating a targeted set of hard negative samples, for example, patches dominated by atmospheric phase gradients or located in layover and shadow regions, or by conducting an ablation study comparing different positive-to-negative sampling ratios and their effect on the false positive rate. This would provide important practical guidance for users seeking to deploy the model operationally over large areas.

We thank the reviewer for this suggestion and acknowledge that incorporating difficult background signals may help to better control false positives.

We would like to clarify that due to the pixel-wise nature of the CNN, the patch size (128 × 128 pixels; ~1.6 km × 1.6 km) and our "positive-only" dataset being defined to have at least one pixel to be annotated as mass movement (see lines 221-23), the vast majority of training pixels (about 80% of each sample patch) are already negative. This means, even without explicitly including purely negative patches, the model is exposed to substantial background variability within positive patches, including atmospheric artefacts, decorrelation noise, and geometric distortions.

We have so far refrained from explicitly including negative patches in the final training, as

i) 78% or more of each sample patch is already acting as negative training data

ii) initial tests using an active learning approach including false positives (hard negative mining) indicated only minor improvements

iii) a test using naïve random sampling of negative patches yielded no measurable improvement, as most negative patches could fall on easy to segment areas, and

iv) we cannot guarantee that targeted candidate negative patches especially in the complex terrain are truly free of deformation signals. Given the known underrepresentation of smaller or uncertain mass movements, some "hard negatives" may in fact correspond to unmapped or ambiguous signals, thereby introducing label noise and potentially confusing the model.

v) the IoU change between positive only patches and the full dataset is ~0.04, which indicates the model can robustly segment negative patches

- The authors identify the classification of phase patterns as coherent versus incoherent as the primary driver of inter-expert disagreement, with mean IoU values rising substantially from 0.21–0.41 to 0.496–0.644 when this distinction is not enforced (Section 3.1). This finding is significant, yet its implications for the DL training design are not fully explored in the discussion. Specifically, because the boundary between coherent and incoherent signals is itself ambiguous among experts, as demonstrated empirically by the study, the decision to train the model exclusively on coherent phase patterns may introduce compounding label uncertainty at the class boundary and could constrain the model's ability to generalize across the full spectrum of mass movement signatures encountered in operational interferograms. The authors are encouraged to discuss this connection more explicitly in Section 4.2.1 or 4.2.3, and to clarify how the observed ambiguity in coherent versus incoherent classification informed the DL model's scope and how users should interpret predictions near this classification boundary when applying the model in practice.

We thank the reviewer for catching this important point. In the revised manuscript we will clarify why we deliberately chose to focus the DL training exclusively on coherent signals even though the distinction between coherent and incoherent phase patterns was identified as the primary driver of inter-expert disagreement.

This decision was motivated by the substantially different nature of incoherent signals. From a data perspective, the detection of incoherent areas represents a different task, as these signals differ fundamentally in appearance from coherent deformation patterns and are more easily confused with coherence losses unrelated to mass movements (e.g., vegetation or water bodies). Experts rely on information beyond a single interferogram to interpret incoherent areas and to attribute the loss of coherence to a specific fast moving process. As a result, including incoherent patterns would introduce additional sources of ambiguity that are difficult to constrain within the DL training framework since it would complicate the clear definition of a training goal.

We therefore adopted a pragmatic approach and defined a more specific modelling objective: the detection of coherent mass-movement signals. In the revised discussion (Sections 4.2.1 and 4.2.3), we will make this design choice more explicit and acknowledge that the ambiguity in expert classification near the coherence boundary propagates into the training labels and may affect model predictions in this transition zone. Accordingly, we will frame the model more clearly as primarily detecting coherent mass-movement signals and emphasise that predictions close to the coherence boundary should be interpreted with caution in operational applications.

- Section 2.1 describes the D-InSAR processing workflow using Sentinel-1 data acquired in ascending (track A088) and descending (track D066) geometries. While the authors have appropriately leveraged both viewing geometries for training data generation, there is a growing body of literature demonstrating that three-dimensional surface displacement fields can be retrieved by fusing multiple line-of-sight InSAR datasets from complementary acquisition geometries. Such multi-looking approaches, combining ascending and descending passes, and in some cases integrating azimuth offset tracking,

> can provide substantially richer kinematic information about mass movement processes, including the decomposition of horizontal and vertical displacement components that is often critical for interpreting failure mechanisms and movement styles. Given that the study already processes both ascending and descending interferograms, a brief discussion of how three-dimensional displacement retrieval could complement or extend the proposed DL detection framework would strengthen the contextual framing of the work and better position it within the current state of the art. The authors are encouraged to at minimum acknowledge these capabilities and their relevance to the broader landslide monitoring workflow in the introduction or discussion. Relevant examples from the recent literature include:
> doi:10.3390/rs11030241; doi.org/10.1016/j.geoai.2026.100061;
> doi:10.1016/j.earscirev.2014.02.005

Thanks for this interesting comment!

The aim of the presented work is to develop a method for efficiently extracting information directly from interferograms in order to detect and delineate mass-movement signals over large areas with minimal processing requirements. The approaches mentioned by the reviewer requires several additional processing steps, including phase unwrapping to derive line-of-sight displacement for each track and the subsequent inversion to estimate three-dimensional displacement components. These steps form part of a later stage in the exploitation of InSAR data products, introduce additional sources of uncertainty, and substantially increase computational complexity. We note that 3-dimensional reconstruction based on InSAR only would be possible only by making strong assumptions on the along track displacement (ca. north-south), as this component is barely measurable considering satellites relying on polar orbits as Sentinel-1 (Manconi et al., 2024). As our proposed deep-learning framework operates directly on wrapped interferograms and does not rely on displacement retrieval, the integration of three-dimensional displacement reconstruction methods does not have a direct methodological link to the approach presented here. For this reason, we consider these developments to be outside the scope of the present paper and have therefore deliberately not included a discussion of them in the manuscript.